LEARNING SPSS *without* PAIN

A Comprehensive Guide for Data Analysis and Interpretation of Outputs

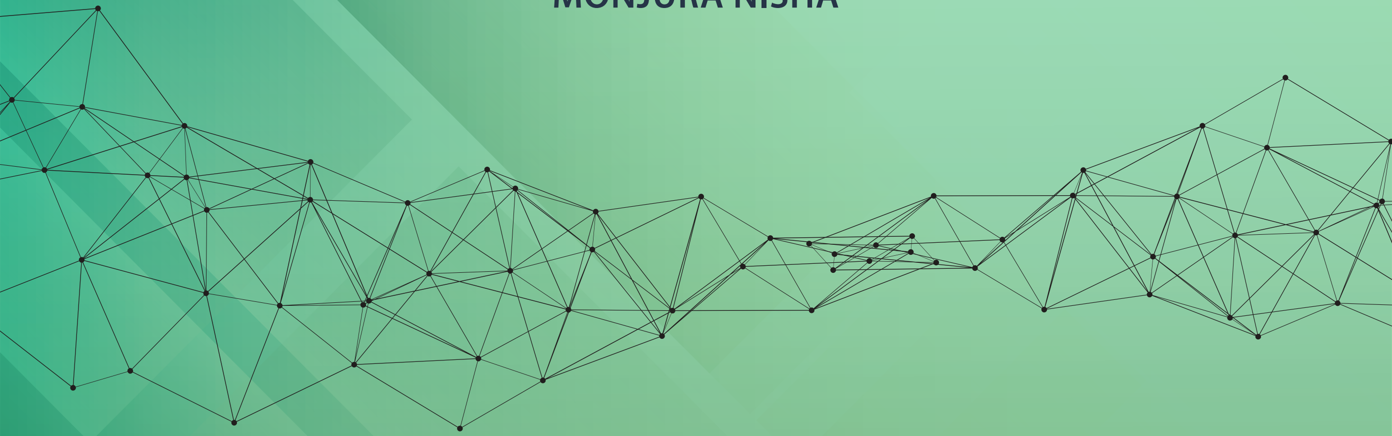
Second Edition

SPSS Version 25

MOHAMMAD TAJUL ISLAM

RUSSELL KABIR

MONJURA NISHA



Learning SPSS without Pain

A Comprehensive Guide for Data Analysis and Interpretation of Outputs

Second Edition

SPSS Version 25

Mohammad Tajul Islam
MBBS, DTM&H, MSc (CTM), MPH
Professor (Adjunct), Department of Public Health
North South University and State University of Bangladesh

Russell Kabir
BDS, MPH, MSc, PhD
Senior Lecturer (Research Methods), Course Leader (MSc Public Health)
Anglia Ruskin University, UK

Monjura Nisha
BDS, MPH, PhD
Sydney School of Public Health
The University of Sydney, Australia

ASA Publications
Dhaka, Bangladesh

ISBN 978-984-34-8254-9 [First Edition]
ISBN 978-984-35-0870-6 [Second Edition]

First published 2020

Publisher
ASA Publications
Dhaka, Bangladesh
Email: asapublications7@gmail.com

Distributor
Altaf Medical Book Center
Shop: 121 & 128; Lane: 3; Islamia Market, Nilkhet, Dhaka 1205
Cell: +880-1711-985991; +880-1611-985991; +880-1511-985991

Production credits
Editor: Mohammad Tajul Islam
Proofreading: Khrieu Healy
Composition and Cover Design: Zariath Al-Mamun Badhon
Production Director: Md Altaf Hossain
Printing: ASA Publications
Binding: Rahim Bindings

Printed in Bangladesh

To all our family members, students and promising researchers in health and social sciences

Preface to second edition

This book is based on SPSS version 25, while the first edition was based on version 16. Many students and researchers still use the lower versions, which are efficient enough to analyze data for the commonly used statistical methods. However, the new generation of students and researchers are inclined to use the recent versions. The newer versions usually accommodate more advanced and recently developed techniques.

This book is intended to prepare the students and promising researchers for the challenging and rewarding job of data management and analysis. The success of this book has encouraged us to publish the second edition. In revising the book, we have considered our careful critical reviews and valuable suggestions from the users. The objective and target group for this book remains the same. The target groups include the students (MPH, MSc, MD, FCPS, MS, MPhil and PhD) of health and social sciences, aspiring young researchers and teachers who have some skills in data analysis and want to become more effective. The users are encouraged to apply the right statistical methods taking the guidance from this book for data management and analysis to become more competent as a person and as a data analyst.

This edition is more comprehensive than the previous edition. While we have retained the topics that are essential and were well received, some chapters have been reorganized and some new topics have been added. The number of chapters has been reduced from 22 in the first edition to 20 in this edition. The new topics and features added in this edition include:

- New illustrations
- Conditional logistic regression
- Multinominal logistic regression
- Cox regression with constant time
- Some chapters have been reorganization for better understanding

With the inclusion of more illustrations and topics, we believe the users will find it easier to analyze their data and interpret the outputs. If you have any comments about the book, feel free to write at the e-mail address below.

M. Tajul Islam
abc.taj@gmail.com

Preface to first edition

This book is intended for the students (MPH, FCPS, MD, MS, MPhil, and others), teachers and young researchers in health and social sciences. It is written in a very simple language and used health related data as examples. This book answers three basic questions related to data analysis. They are: a) what to do (what statistics to be used for the data analysis to achieve the objectives); b) how to do (how to analyze data by SPSS); and c) what do the outputs mean (how to interpret the outputs)? All these questions are answered in a very simple and understandable manner with examples. This book covers the basic statistical methods of data analysis used in health and social science research. It is the gateway to learn SPSS and will help the users to go further. This book is organized in 22 sections that covers data management, descriptive statistics, hypothesis testing using bivariate and multivariable analysis and others. It is easier to learn through exploration rather than reading. The users can explore further once the basics are known. From my understanding, using the statistics as covered in this book, the students and researchers will be able to analyze most of their data on epidemiological studies and publish them in international peer-reviewed journals.

I am optimistic that this book will make the students' and researchers' life easier for analyzing data and interpreting the outputs meaningfully. If you have any comments about the book, feel free to write at the e-mail address below.

M. Tajul Islam
abc.taj@gmail.com

Foreword

I was delighted to be asked to write a foreword to this book. The authors have given much thought to creating a user-friendly textbook, which goes a long way towards turning the 'pain' of statistics into 'gain'. The aim of the book is to make SPSS navigation easy and accessible. No doubt it will become a handy companion to students and researchers in academia and industry. With its step by step visuals and hypothetical yet relevant dataset examples, the handbook provides 'start to finish' guidance from coding variables to inferential analyses.

Since my undergraduate days almost two decades ago now, SPSS has been a go to tool for my own research and teaching as a psychologist and academic. At university, I think it is fair to say that students have a 'love/hate' relationship with SPSS. It is the tool for our projects and often dissertation outputs in the social sciences. Yet, whilst it can yield exciting answers and future research questions, we can feel apprehensive about data coding and choosing the 'right' statistical tests. This textbook guides us through these very steps, reducing room for error whilst helping to build confidence through practice.

The authors of the textbook are both long standing experts in health service and social science research methods - the reader is in good hands. It is worth having a look at some of their academic papers utilizing population-level data to understand and improve public health provision globally - many of these research outputs have come to fruition with the aid of SPSS.

The book is a comprehensive guide for data analysis and interpretation. However, we cannot ignore how overwhelmed some of us can feel when faced with the prospect of statistics. The book goes some way towards easing this frustration. Wherein lies its root? What is it about statistics that leads some of us to experience almost visceral pain? Perhaps, they take us back to school and competitive learning approaches to mathematics. Maybe, they feel complex, with much to learn and even more to compute.

For me, as a qualitative researcher, and now a Trainee Person-Centered Psychotherapist, words, feelings, beliefs have always spoken to me more directly. There is something about numbers that has at times felt reductionist and removed from immediate human experience. Yet, it is through statistics that we can generalize our findings, plan, understand relationships between variables and make predictions and decisions. No doubt, the satisfaction of an SPSS output which addresses our hypotheses, and shapes our understanding of reality can feel like our very own 'eureka' moment. Within the social sciences, this has large-scale implications for resource allocation, and in turn, evidenced-based policy-making and human life quality.

As a parting word from me, the authors have not been paid for the painstaking hours it took to them compile this book. Instead, they ask for a donation to their chosen cause. If it took to them compile this book. Instead, they ask for a donation to their chosen cause. If you can, please demonstrate your appreciation of their efforts by giving generously.

Dr. Maria Kordowicz, FRSA
Chartered Psychologist and Head of Learning, Research and Evaluation
The Social Innovation Partnership
United Kingdom

Acknowledgements

We are particularly indebted and grateful to Dr. Harun Or Rashid and Dr. Shakil Ahmed, who reviewed all the sections and provided their critical views and constructive suggestions. Dr. Shakil helped us in editing the final manuscript and took the lead role in publishing this book. We would like to thank Khrieu Healy from Anglia Ruskin University for proofreading the book for us. We are particularly thankful to Professor Hafiz T. A. Khan of the University of West London, and Dr. Ali Parsa and Dr. Divya Vinnakota of Anglia Ruskin University for their careful review of this book and useful suggestions.

Many of my students, who continually pushed and encouraged me to write this book deserve a share of the credit, including Dr. SM Anwar Sadat, Dr. Md. Naymul Hasan, Md. Golam Kibria, Dr. Sinthia Mazumder, Humayun Kabir, Dr. Rubana Islam and many others. Finally, we express our sincerest thanks for the assistance that we have received from the ASA Publications for publishing the book. Being the authors, we accept full responsibility for any deficiencies that the book may have.

Links for the e-book and datasets:

The users can download the e-book and hypothetical datasets used in this book from the links below.

- https://github.com/rubyrider/Learning-SPSS-without-Pain
- https://drive.google.com/drive/folders/1rfkik8iHho0bM5QoQMQ JTDbCKedWhd2C?usp=sharing

To the users

Of course, this e-book is free for all. However, if you can afford, please donate Tk.100 only (US$ 2 for the users outside Bangladesh) to any charity or a needy person. This little amount is sufficient to offer a meal for an orphan in developing countries.

Contents

1

Introduction

SPSS stands for Statistical Package for Social Sciences. It is a powerful window-based statistical data analysis software. The menu and dialog-box system of SPSS have made the program user-friendly. SPSS is particularly useful to the researchers in public health, medicine, social science and other disciplines. It supports a wide range of univariate, bivariate, multivariable and multivariate data analysis procedures.

This book is intended for the students, teachers and researchers involved in health and social sciences research. It provides practical guidance of using SPSS for basic statistical analysis of data. Once the data file is loaded in SPSS, the users can select items from a dropdown menu to analyze data, make graphs, transform variables, and others. SPSS, in general, has made the life of the researchers easier for data analysis.

This book is based on SPSS Version 25. The SPSS version 25 is mostly similar to other versions. For the commonly used statistical methods of data analysis, the commands of SPSS Version 25 are almost the same as those of other versions, though there are some differences in commands for some statistical analyses, when compared with the lower versions (especially the Versions 19 and below). The newer versions are usually for the use of more advanced and recently developed techniques. Those who have access to the lower versions of SPSS (especially version 19 and below) can still use this book or the first edition of this book.

This book is primarily developed targeting the Master of Public Health (MPH) and post-graduate students in medicine (FCPS, MD, MS, and MPhil), keeping in mind their needs. The aim of this book is to provide a concise but clear understanding on how to conduct a range of statistical analyses using SPSS and interpret the outputs. Special emphasis is given to understanding the SPSS outputs, which is a problem for many of the users. For a better understanding of the users, examples and data related to health research are used. However, to use the book effectively and to understand the outputs, it requires basic knowledge on biostatistics and epidemiology. The users will find it easier if they review the relevant statistical concepts and procedures, and epidemiological methods before using this book.

1.1 Steps of data analysis

We collect data for our studies using various tools and methods. The commonly used tools for data collection are questionnaires and record sheets, while the commonly used data collection methods are face-to-face interviews, observations, physical examinations and lab tests. Sometimes we use the available data (secondary data) for our research studies, for example, hospital records, and data of other studies (e.g., Bangladesh Demographic and Health Survey data). Once data is collected, the steps of data analysis are:

- Data coding, if pre-coded questionnaire or record sheet is not used
- Development of data file and data entry
- Data cleaning (checking for errors in data entry)
- Data screening (checking assumptions for statistical tests)
- Data analysis
- Interpretation of results

In the following sections, we have discussed the development of data files, data management, data analysis and interpretation of the outputs.

2

Generating Data Files

Like other data analysis programs, SPSS has to read a data file to analyse data. We, therefore, need to develop a data file for the use of SPSS. The data file can be generated by SPSS itself or by any other program. Data files generated in other programs can be easily transferred to SPSS for analysis. Here, we will discuss how to generate a data file in SPSS.

2.1 Generating data files

The first step in generating a data file is to give a "name" and "define" variables included in the questionnaire/ record sheet. The next step is entering data in SPSS. Suppose we have collected data using a pre-coded questionnaire (codes are shown in the parenthesis) with the following variables:

Categorical variables:

- Sex (m= male; f = female)
- Religion (1= Islam/Muslim; 2= Hindu; 3= Others)
- Occupation (1= Business; 2= Government job; 3= Private job; 4= Others)
- Marital status (1= Married; 2= Unmarried; 3= Others)
- Have diabetes mellitus (1= Yes; 2= No; 3= Don't know)

Quantitative variables (numerical variables):

- Age of the respondent
- Monthly family income
- Systolic blood pressure (BP)
- Diastolic BP

Suppose we have decided to use V1 as the SPSS variable name for age, V2 for sex and V3 for religion (Table 2.1). Instead of V1, V2, V3, you can use any other variable name (e.g., age for age and sex for sex) for your variables. It is always better to develop a code-

-book in MS Word or MS Excel before entering data, as shown in Table 2.1. This is helpful during data analysis.

Table 2.1 Codebook

SPSS variable name	Actual variable name	Variable code
V1	Age in years	Actual value

V2	Sex	m= Male
		f= Female

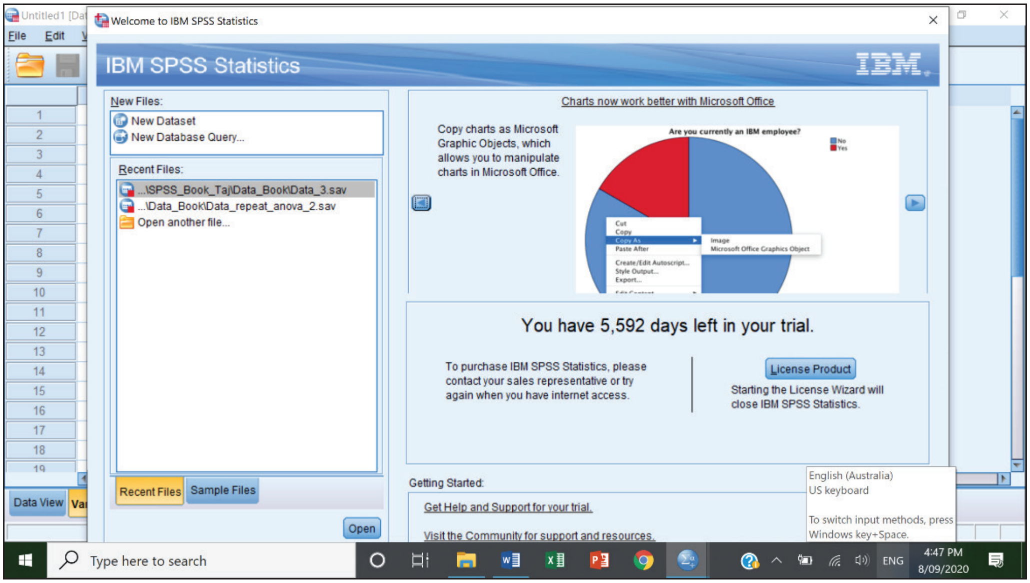
V3	Religion	1= Islam/Muslim 2= Hindu 3= Others
V4	Occupation	1= Business 2= Government job 3= Private job 4= Others
V5	Monthly family income in Tk.	Actual value
V6	Marital status	1= Married 2= Unmarried 3= Others
V7	Have diabetes mellitus	1= Yes 2= No 3= Don't know
V8_a	Systolic blood pressure in mmHg	Actual value
V8_b	Diastolic blood pressure in mmHg	Actual value

Note: *Instead of V1, V2, you can use any other name as SPSS variable name. For example, you can use the variable name "age" instead of V1 and "sex" instead of V2.*

Open the SPSS program by double-clicking the SPSS icon. You will see the following dialogue box (Fig 2.1). Click on cancel box (×) to close the "IBM SPSS Statistics" window. Now we have the dialogue box as shown in Figure 2.2 (IBM SPSS Statistics Data Editor). This is the window for defining the variables.

The "IBM SPSS Statistics Data Editor (Fig 2.2)" shows Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure, and Role at the top row. If you do not see this, click on "Variable View" at the left-bottom corner of the window.

Figure 2.1 IBM SPSS opening dialogue box


Untitled1 [Da
File Edit
Welcome to IBM SPSS Statistics
IBM SPSS Statistics
IBM.
New Files:
New Dataset
New Database Query...
Recent Files:
...\SPSS_Book_Taj\Data_Book\Data_3.sav
...\Data_Book\Data_repeat_anova_2.sav
Open another file...
Recent Files
Sample Files
Open
Charts now work better with Microsoft Office
Are you currently an IBM employee?
No
Yes
Cut
Copy
Copy As
Paste After
Create/Edit Autoscript...
Style Output...
Export...
Image
Microsoft Office Graphics Object
License Product
Getting Started:
Get Help and Support for your trial.
Visit the Community for support and resources.
English (Australia)
US keyboard
Data View
Va
Type here to search
ENG
4:47 PM
8/09/2020

Copy charts as Microsoft Graphic Objects, which allows you to manipulate charts in Microsoft Office.

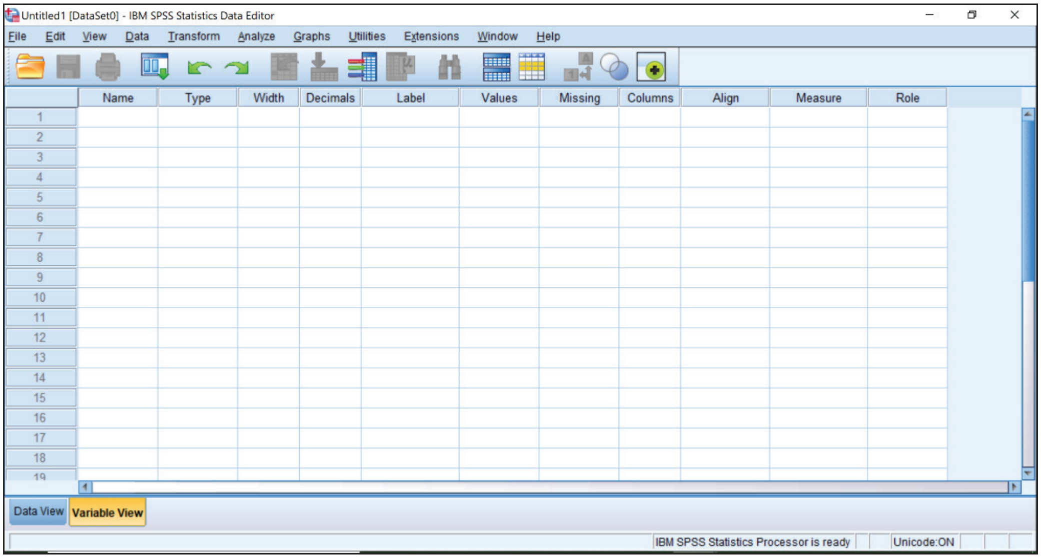
You have 5,592 days left in your trial.

To purchase IBM SPSS Statistics, please contact your sales representative or try again when you have internet access.

Starting the License Wizard will close IBM SPSS Statistics.

To switch input methods, press Windows key+Space.

Figure 2.2 IBM SPSS statistics data editor: template for defining variables


Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Name
Type
Width
Decimals
Label
Values
Missing
Columns
Align
Measure
Role
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Data View
Variable View
Unicode:ON

IBM SPSS Statistics Processor is ready

2.1.1 Defining variables

We shall use the IBM SPSS statistics data editor (Fig 2.2) to define the variables. Before entering data, all the study variables need to be defined including their coding information. The lower versions of SPSS (version 12 and below) allow only 8 characters to name a variable. The higher versions (13 and above) allow up to 64 characters to name a variable. While writing the variable names, we need to follow certain rules. They are:

- The variables must be unique (all variables should have different names)
- Variables must begin with a letter (small or capital) rather than a number
- Cannot include full stop (.), space or symbols, like ?, *, µ and λ
- Cannot include words that are used as commands by SPSS, such as ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO and WITH
- Cannot exceed 64 characters

To define the variables, follow the instructions below.

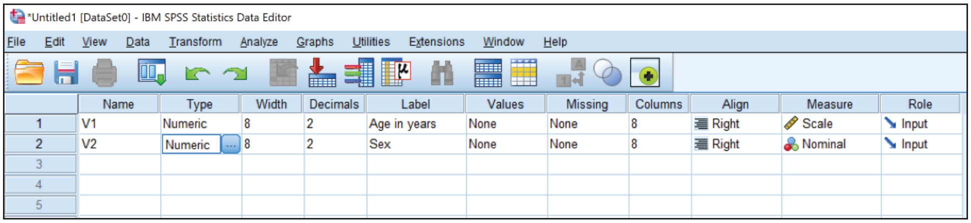
Name: The "Name" column is for writing the brief SPSS variable names, as shown in the codebook. Click on the first cell of the "Name" column. Type the brief SPSS variable name of the first variable in this cell. For example, type V1 (you can also use "age" as the variable name) for the first variable, age. Note that this short name will be used to identify the variable in the data file.

Type: This column indicates the characteristic of the variable, whether it is a numeric or string variable. The numeric variables are expressed by numbers (e.g., 1, 2, 3), while the string variables are expressed by alphabets (e.g., m, f, y, n). In SPSS, the default value for "Type" is numeric. If the nature of the variable is string (alphabet or text variable), we need to change it. To change the variable type into string (or other types), use the following steps:

- Click on the cell under the column "Type" (you will see a box with three dots; Fig 2.3)
- Click on the "three-dot box" (you will see the options in a separate dialogue box; Fig 2.4)
- Select "String" from the options, then click Ok

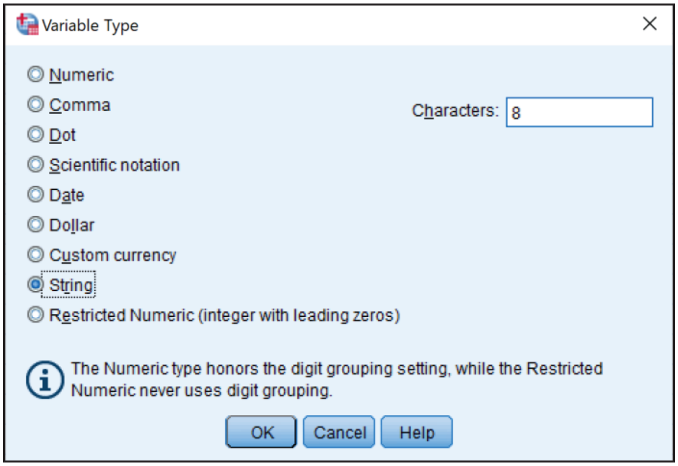
Similarly, if it is a date variable (e.g., date of hospital admission), you have to change the variable type into a date format in the same manner.

Figure 2.3 Template for defining variables



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	V1	Numeric	8	2	Age in years	None	None	8	Right	Scale	Input
2	V2	Numeric ...	8	2	Sex	None	None	8	Right	Nominal	Input
3											
4											
5											

Figure 2.4 Template for selecting variable type



○ Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
● String
○ Restricted Numeric (integer with leading zeros)

Characters: 8

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Width: The default value for width is 8. In most cases, it is sufficient and serves the purpose. However, if the variable has very large value, then we need to increase it using the arrow (up and down) button in the box. *For practical purposes, keep the width 8 unless the variable values are larger than 8 characters.*

Decimals: This is applicable for the numeric variables, and the default value is 2. If your data does not have any decimal value, you can make it "0" using the down arrow or keep it as it is.

Label: This is the space where we write the longer description of the variable (actual variable name, as shown in the codebook). For example, we have used "V1" to indicate age in years. We should, therefore, write "Age in years" in the label column for the variable name "V1".

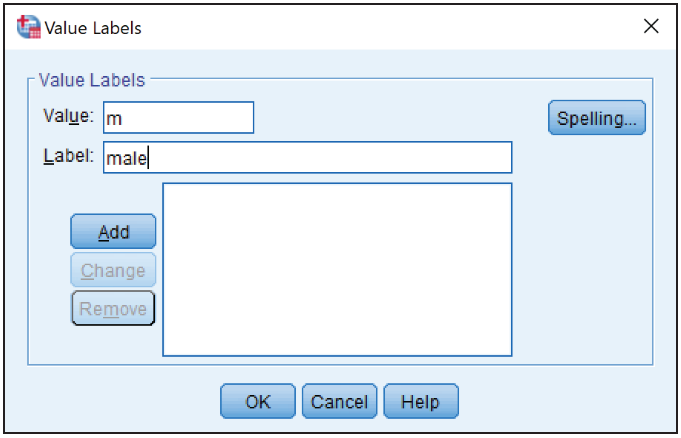
Values: This is applicable for the variables to define their levels (categories) using code numbers (such as 1, 2 or m, f). This allows the SPSS to retain the meaning of values (code numbers) you have used in the dataset. For example, our variable 2 is 'sex' and is defined by "V2". It has two levels/categories, male (coded as "m") and female (coded as "f"). Follow the commands below to put the value labels.

- Click on cell under the column "Value" (you will see a box with three dots)
- Click on "three-dot box"
- Click in the box "Value" and Type "m"
- Click in the box "Value label" and Type "male" (Fig 2.5)
- Click "Add"

- Repeat the same process for female (value “f”, value label “female”, add) and then click OK

Follow the above steps to complete the value labels of all the variables, if applicable. *Note that value labels are needed only for the variables that have been coded.*

Figure 2.5 Template for value labels



Value: m

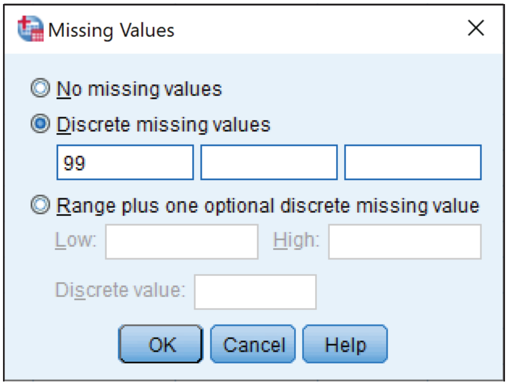
Label: male

Missing: If there is any missing value in the dataset, SPSS has the option to indicate that. To set a missing value for a variable, we must select a value that is not possible (out of range) for that variable. For example, we have conducted a study and the study population was women aged 15-49 years. There are several missing values for age in the data (i.e., age was not recorded on the questionnaire or respondent did not tell the age). First, we need to select a missing value for age. We can select any value which is outside the range 15-49 as the missing value. Let’s say we decided to use 99 as the missing value for age. Now, to put the missing value for age in SPSS, use the following commands.

Click on the cell under the column “Missing” (you will see a box with three dots) > Click on the “three-dot box” > Select “Discrete missing values” (Fig 2.6) > Click in the left box > type “99” > Click OK

However, you may omit this. Just keep the cell blank while entering data in the data file. SPSS will consider the blank cells in the data file as missing values (system missing).

Figure 2.6 Missing value template



Missing Values

○ No missing values

● Discrete missing values

99

○ Range plus one optional discrete missing value

Low: High:

Discrete value:

Columns: The default value for this is 8, which is sufficient for most of the cases. If you have a long variable name, then only change it as needed. For practical purposes, just keep it as it is.

Align: You do not need to do anything about this.

Measure: This cell indicates the measurement scale of the data. If the variable is categorical, use "Nominal" for the nominal variable or "Ordinal" for the ordinal variable. Otherwise use "Scale" for interval or ratio scale of measurement.

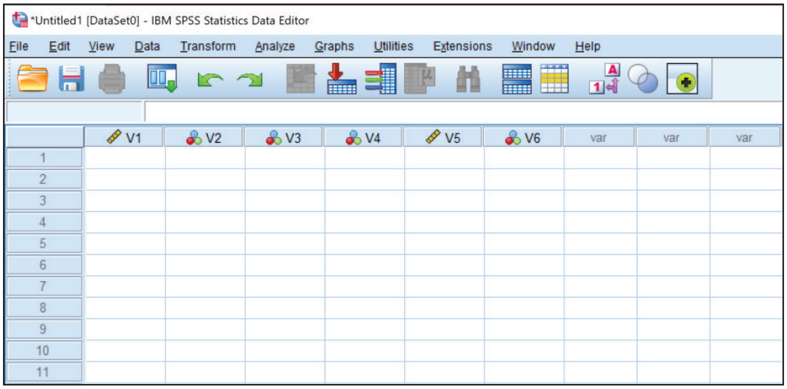
Role: Some dialogues support predefined roles that can be used to pre-select variables for analysis. The default role is "Input". You don't need to do anything, just keep it as it is.

In this way, define all the variables of your questionnaire/record sheet in the SPSS data editor. The next step is data entry.

2.1.2 Data entry in SPSS

Once all the variables are defined, click on the "Data View" tab at the bottom-left corner of the window. You will see the template with the variable names at the top row (Fig 2.7). This is the spreadsheet for data entry. Now you can enter data starting from the first cell of row 1 for each of the variables. Complete your data entry in this spreadsheet and save the data file at your desired location/folder (save the file as you save your file in MS Word, such as click on File> Click on Save as > Select the desired folder > Give a file name > Save).

Figure 2.7 Spreadsheet for data entry (SPSS statistics data editor)



If you want to open the data file later, use the following commands.

Click on File > Open > Data > Select the folder you have saved your SPSS data file > Select the file > Click "Open"

2.2 Data used in this book

The following data files have been used in this book as examples, which are available at the following links. Readers are welcome to download the data files for practice. The data files (with hypothetical data) used in this book include:

- Data_3.sav
- Data_4.sav
- Data_5 multinominal.sav
- Data_Ca-Co_Matched.sav
- Data_HIV.sav
- Data_repeat_anova_2.sav
- Data_survival_4.sav
- Data_cronb.sav

Links for the data files:

You can download the e-book and data files from any of the links below.

Link 1: https://github.com/rubyrider/Learning-SPSS-without-Pain

Link 2: https://jmp.sh/F65fcni

Link 3: https://drive.google.com/drive/folders/1rfkik8iHho0bM5QoQMQ-JTDbCKedWhd2C?usp=sharing

3

Data Cleaning and Data Screening

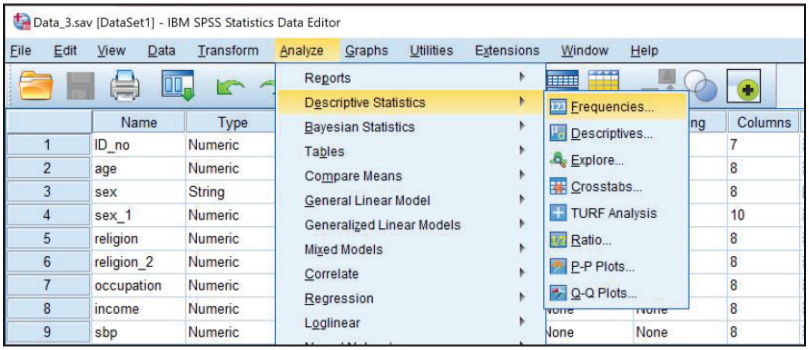
Once data has been entered into SPSS, we need to be sure that there are no errors in the data file (i.e., there were no errors during data entry). Data cleaning is commonly done by generating frequency distribution tables of all the variables to see the out-of-range values, and by cross tabulations (or by other means) for checking the conditional values. If errors are identified, they need to be corrected. Simultaneously, we also need to check the data if it fulfils the assumptions of the desired statistical test (data screening), e.g., is data normally distributed to do a t-test? *The users may skip this chapter for the time being and go to chapter 4.* Once the users develop some skills in data analysis, they can come back to this chapter. Use the data file <**Data_3.sav**> for practice. The codebook of this data file can be seen in the annex (Table A.1).

3.1 Checking for out-of-range errors

We can check the out-of-range errors by making a frequency distribution of the variable or using the statistics option for minimum and maximum values. For example, you want to check if there are any out-of-range errors in the variable “religion” (note that the variable “religion” has 3 levels/values: 1= Islam/Muslim; 2= Hindu; 3= Others). To check for out-of-range errors, we shall find the minimum and maximum values of religion in the dataset using the following commands.

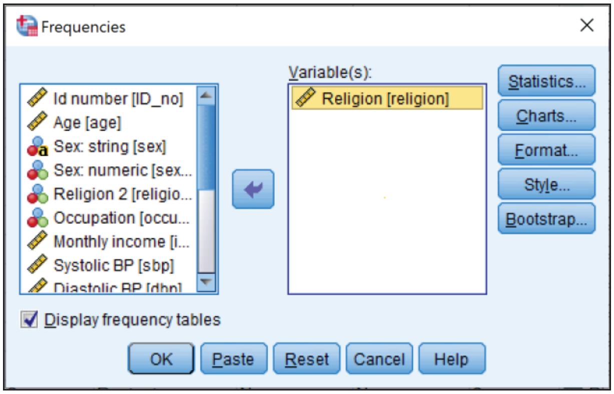
> Analyze > Descriptive statistics > Frequencies > Select the variable “religion” and push it into the “Variable(s)” box > Statistics > Select “minimum” and “maximum” under “Dispersion” section > Continue > OK (Fig 3.1 to 3.3)
>
> With these commands SPSS will produce Table 3.1.

Figure 3.1



	Name	Type	Columns
1	ID_no	Numeric	7
2	age	Numeric	8
3	sex	String	8
4	sex_1	Numeric	10
5	religion	Numeric	8
6	religion_2	Numeric	8
7	occupation	Numeric	8
8	income	Numeric	8
9	sbp	Numeric	8

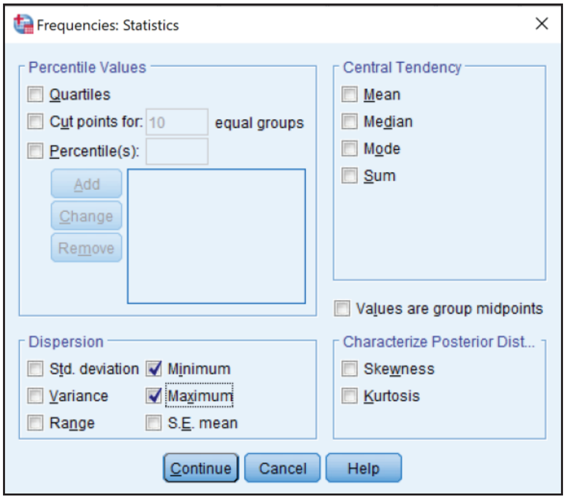
Figure 3.2



Variable(s): Religion [religion]

☑ Display frequency tables

Figure 3.3



Percentile Values

□ Quartiles

□ Cut points for: 10 equal groups

□ Percentile(s):

Central Tendency

□ Mean

□ Median

□ Mode

□ Sum

□ Values are group midpoints

Dispersion

□ Std. deviation ☑ Minimum

□ Variance ☑ Maximum

□ Range □ S.E. mean

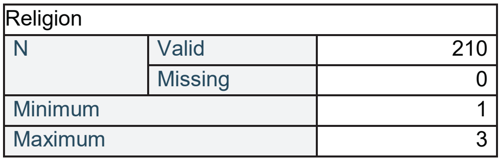
Characterize Posterior Dist...

□ Skewness

□ Kurtosis

Look at Table 3.1 (first Table) of SPSS output. You can see if there is any value which is out of the range 1-3 in the dataset in the table. If there is any value >3 or <1, this means that there is an error in data entry. In this scenario, identify the subject (by ID no.) and correct the error by referring to the filled out questionnaire.

Table 3.1 Statistics



Religion		
N	Valid	210
	Missing	0
Minimum		1
Maximum		3

Table 3.1 shows that the values range from 1 to 3 (minimum 1 and maximum 3), which are within the range of our code numbers. Therefore, there is no out-of-range error in this variable.

3.2 Checking for outliers

Outliers are extreme values that deviate from other observations. The outliers may appear in a dataset because of errors or variability in the measurements, errors in data entry, or other reasons. A commonly used rule says that a data point is an outlier if it is more than 1.5 × IQR (inter-quartile range) above the 3rd quartile (i.e., Q3 + 1.5 × IQR) or below the 1st quartile (i.e., Q1 – 1.5 × IQR). There are several ways to identify the outliers in a dataset. The most commonly used method to detect outliers is visualization of data. Examples of visualization methods that can be used are box-plot chart, scatter plot and histogram. Statistical methods like Z-score and other methods are also used.

We shall check the potential outliers by constructing the box and plot chart. Outliers are indicated by ID numbers on the chart. Outliers are greater than 1.5 box length distance from the edge (upper or lower) of the box. The extreme values are indicated by "*" [greater than 3 box length distance (i.e., 3 × IQR) from the edge of the box]. To construct the box and plot chart for the variable "systolic blood pressure" (SPSS variable name: sbp), use the following commands.

Analyze > Descriptive statistics > Explore (Fig 3.1) > Select "sbp" and push it into the "Dependent list" box > OK

You can also generate the box and plot chart using the following commands.

Graph > Legacy dialogs > Select "Simple" > Select "Summaries of separate variables" under "Data in chart are" > Define > Select "sbp" and push it into the "Boxes represent" box > OK

With the use of the first commands, you will find the box and plot chart (Fig 3.4) along with other outputs (Table 3.2). The Figure 3.4 shows that there are 3 outliers in “systolic blood pressure” as indicated by the ID numbers (20, 54 and 193).

We can also examine the influence of outliers in the data by comparing the 5% trimmed mean (mean of the data after excluding upper 5% and lower 5% of the values) with the mean of the whole dataset. If these two means are close together, there is no influence of outliers in the dataset. Look at Table 3.2. The mean of the systolic blood pressure (BP) is 127.7, while the 5% trimmed mean is 126.5. Since the values are not that different (close to each other), there is no influence of outliers in the data of systolic BP.

Figure 3.4 Box and Plot chart of systolic blood pressure

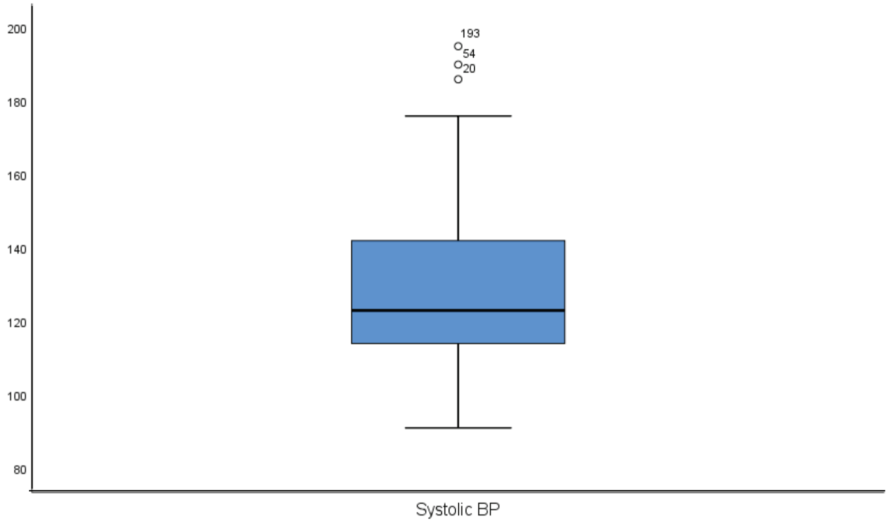


Table 3.2 Descriptive statistics of systolic BP

			Statistic	Std. Error
Systolic BP	Mean		127.73	1.384
	95% Confidence Interval for Mean	Lower Bound	125.00	
		Upper Bound	130.46	
	5% Trimmed Mean		126.58	
	Median		123.00	
	Variance		402.321	
	Std. Deviation		20.058	
	Minimum		91	
	Maximum		195	
	Range		104	
	Interquartile Range		28	
	Skewness		.736	.168
	Kurtosis		.336	.334

3.3 Assessing normality of data

One of the major assumptions for parametric tests is that the dependent quantitative variable is normally distributed. Whether the data has come from a normally distributed population or not, can be checked in different ways. Commonly used methods of checking normality of a dataset are:

- Histogram
- Q-Q plot
- Formal statistical test [Kolmogorov Smirnov (K-S) test or Shapiro Wilk test]

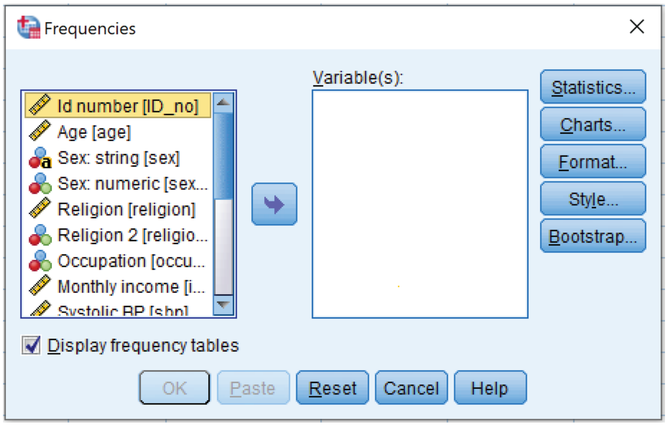
This topic is discussed in detail in chapter 5.

3.4 Variable display in templates

While analyzing data (e.g., frequency distribution or other analysis), SPSS template shows the variable labels along with the short name of the variables in brackets (Fig 3.5). If you want to see only the short name of the variables, use the following commands.

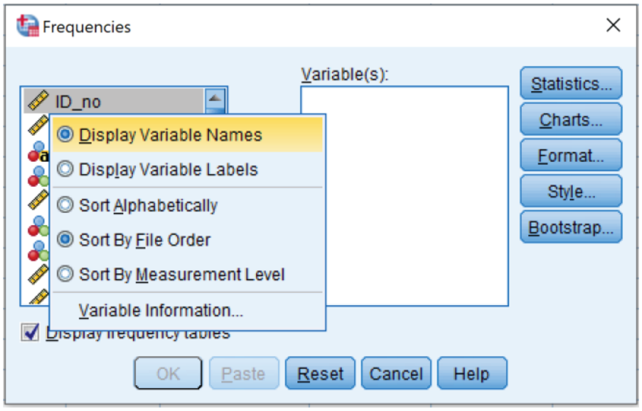
Click on any variable in the variables window (Fig 3.5) > Click right button of the mouse > Select "Display variable names" (Fig 3.6) (there are also other options). Now, you will see only the short variable names in the window (Fig 3.7).

Figure 3.5



☑ Display frequency tables

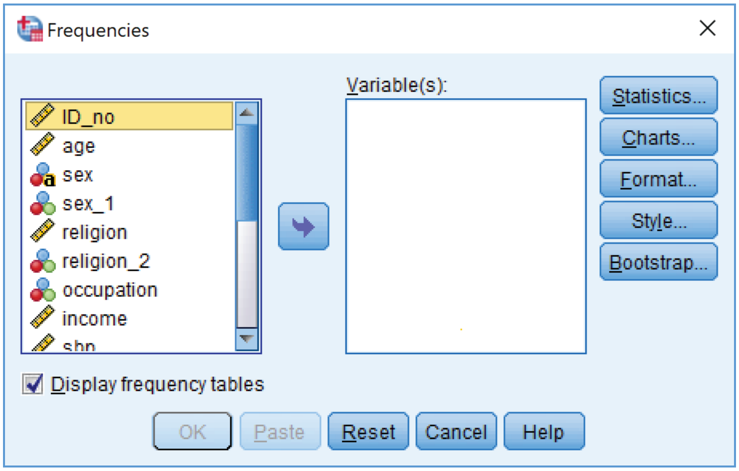
Figure 3.6



● Display Variable Names
○ Display Variable Labels
○ Sort Alphabetically
● Sort By File Order
○ Sort By Measurement Level
Variable Information...

☑ Display frequency tables

Figure 3.7



☑ Display frequency tables

4

Data Analysis: Descriptive Statistics

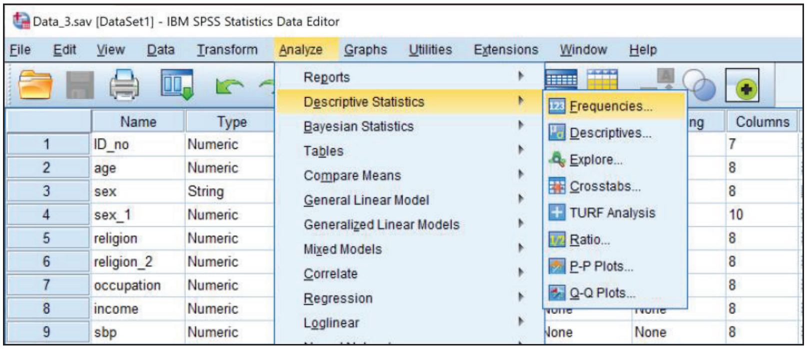
Descriptive statistics are always used at the beginning of data analysis. The objective of using descriptive statistics is to organize and summarize data. Commonly used descriptive statistics are frequency distribution, measures of central tendency (mean, median, and mode) and measures of dispersion (range, standard deviation, and variance). Measures of central tendency convey information about the average value of a dataset, while a measure of dispersion provides information about the amount of variation present in the dataset. Other descriptive statistics include quartile and percentile. Use the data file <**Data_3.sav**> for practice.

4.1 Frequency distribution

Suppose you want to find the frequency distribution of the variables "sex" and "religion". To do this, use the following commands.

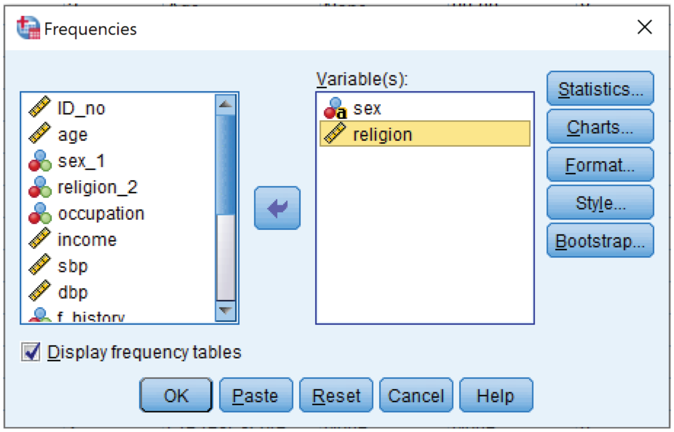
> Analyze > Descriptive Statistics > Frequencies > Select the variables "sex" and "religion" and push them into the "Variable(s)" box > OK (Fig 4.1 and 4.2)

Figure 4.1 Commands for frequency distribution of variables



	Name	Type	Columns
1	ID_no	Numeric	7
2	age	Numeric	8
3	sex	String	8
4	sex_1	Numeric	10
5	religion	Numeric	8
6	religion_2	Numeric	8
7	occupation	Numeric	8
8	income	Numeric	8
9	sbp	Numeric	8

Figure 4.2 Selection of variables for frequency distribution



Variable(s): sex, religion

☑ Display frequency tables

SPSS will provide the following outputs (Table 4.1 and 4.2). Table 4.1 indicates that there are in total 210 subjects, out of which 133 or 63.3% are female and 77 or 36.7% are male. Table 4.2 provides similar information, but for religion.

If there is any missing value in the data, the table will show it (Table 4.3, this is an additional table). In that case, use the *valid percent* instead of *percent* for reporting. For example, Table 4.3 shows that there are 4 missing values in the data. You should, therefore, report 131 or 63.6% are female and 75 or 36.4% are male. *Note that the "Percent" and "Valid Percent" will be the same, if there is no missing value.*

Table 4.1 Frequency distribution of sex with no missing value

Sex: string					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	133	63.3	63.3	63.3
	male	77	36.7	36.7	100.0
	Total	210	100.0	100.0	

Table 4.2 Frequency distribution of religion

Religion					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MUSLIM	126	60.0	60.0	60.0
	HINDU	58	27.6	27.6	87.6
	Christian	26	12.4	12.4	100.0
	Total	210	100.0	100.0	

Table 4.3 Frequency distribution of sex with 4 missing values

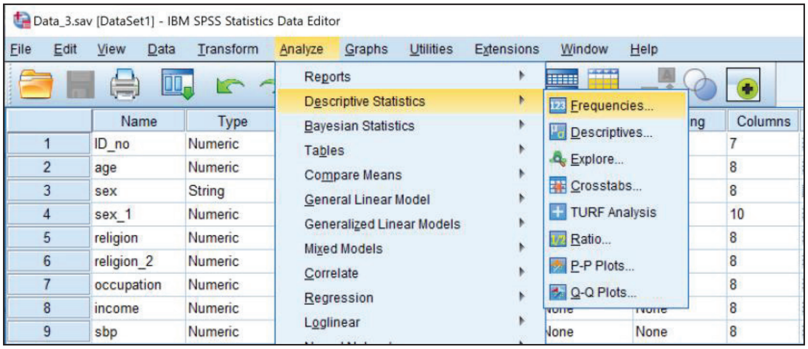
Sex: string					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	131	62.4	63.6	63.6
	male	75	35.7	36.4	100.0
	Total	206	98.1	100.0	
Missing	9	4	1.9		
Total		210	100.0		

4.2 Central tendency and dispersion

We calculate the central tendency and dispersion for the quantitative variables. For example, you want to find the mean, median, mode, standard deviation (SD), variance, standard error (SE), skewness, kurtosis, quartile, percentile (e.g., 30th and 40th percentile), minimum and maximum values of the variable "age" of the study subjects. All these statistics can be obtained in several different ways. However, using the following commands is the easiest way to get those together (Fig 4.3-4.5).

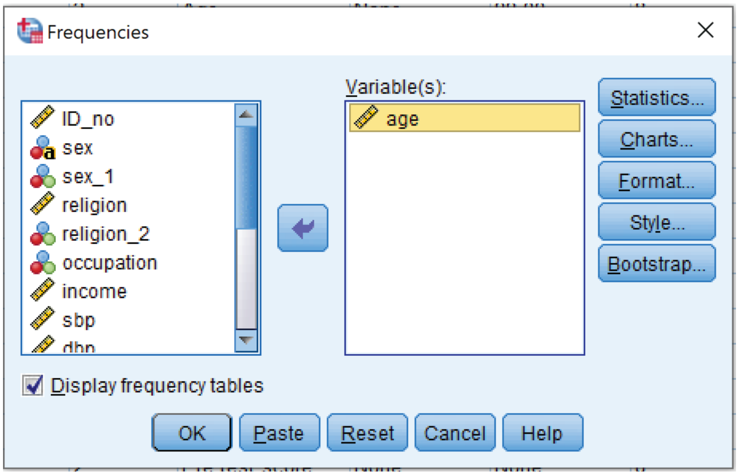
> Analyze > Descriptive statistics > Frequency > Select the variable "age" and push it into the "Variable(s)" box > Statistics > Select all the descriptive measures you desire (mean, median, mode, SD, SE, quartile, skewness, kurtosis) > Select "Percentiles" > Write "30" in the box > Add > Write "40" in the box > Add > Continue > OK

Figure 4.3 Commands for obtaining central tendency and dispersion



	Name	Type	ng	Columns
1	ID_no	Numeric		7
2	age	Numeric		8
3	sex	String		8
4	sex_1	Numeric		10
5	religion	Numeric		8
6	religion_2	Numeric		8
7	occupation	Numeric		8
8	income	Numeric		8
9	sbp	Numeric	None	8

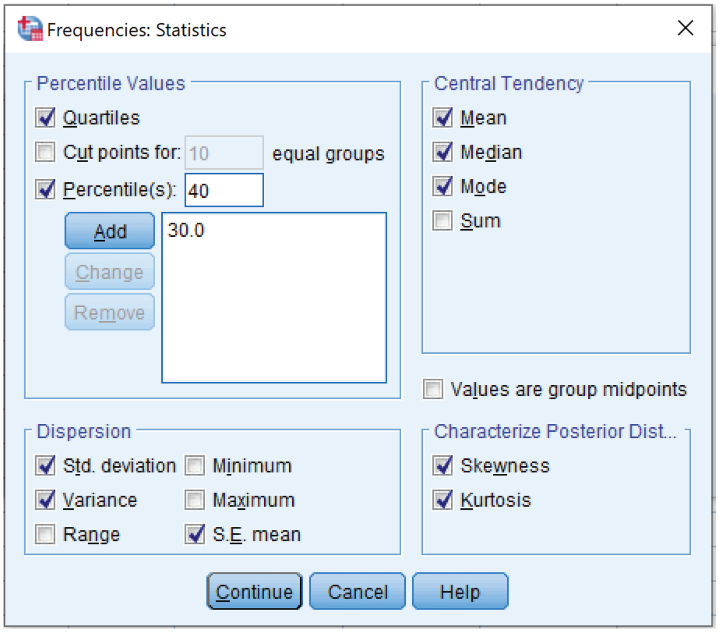
Figure 4.4 Selection of variable(s)



Variable(s): age

☑ Display frequency tables

Figure 4.5 Selection of statistics for the variable(s)



Percentile Values

☑ Quartiles

□ Cut points for: 10 equal groups

☑ Percentile(s): 40

30.0

Central Tendency

☑ Mean

☑ Median

☑ Mode

□ Sum

□ Values are group midpoints

Dispersion

☑ Std. deviation □ Minimum

☑ Variance □ Maximum

□ Range ☑ S.E. mean

Characterize Posterior Dist...

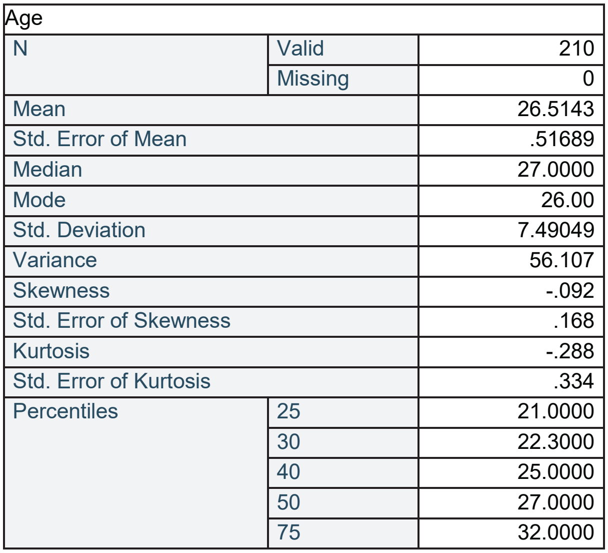
☑ Skewness

☑ Kurtosis

4.2.1 Outputs

SPSS will produce the following output (Table 4.4).

Table 4.4 Descriptive statistics of age



Age		
N	Valid	210
	Missing	0
Mean		26.5143
Std. Error of Mean		.51689
Median		27.0000
Mode		26.00
Std. Deviation		7.49049
Variance		56.107
Skewness		-.092
Std. Error of Skewness		.168
Kurtosis		-.288
Std. Error of Kurtosis		.334
Percentiles	25	21.0000
	30	22.3000
	40	25.0000
	50	27.0000
	75	32.0000

4.2.2 Interpretation

We can see all the descriptive statistics (central tendency and dispersion) that we have selected for the variable "age" including the statistics for Skewness and Kurtosis in Table 4.4. Hopefully, you understand the mean (average), median (middle value of the data set), mode (most frequently occurring value), SD (average difference of individual observation from the mean), variance (square of SD) and SE of the mean. As presented in Table 4.4, the mean age is 26.5 years and SD is 7.49 years. Let me discuss the other statistics provided in Table 4.4, particularly the skewness, kurtosis, quartile and percentile.

Skewness and Kurtosis: These two statistics are used to judge whether the data have come from a normally distributed population or not. In Table 4.4, we can see the statistics for Skewness (- .092) and Kurtosis (- .288). Skewness indicates the spread of the distribution. Skewness ">0" indicates data is skewed to the right; skewness "<0" indicates data is skewed to the left, while skewness "~0" indicates data is symmetrical (normally distributed). The acceptable range for normality of a data set is skewness lying between "-1" and "+1".

However, normality should not be judged based on skewness alone. We need to consider the statistics for kurtosis as well. Kurtosis indicates "peakness" or "flatness" of the distribution. Like skewness, the acceptable range of kurtosis for a normal distribution is between “-1” and “+1”. Data for “age” has skewness -.092 and kurtosis -.288, which are within the normal limits of a normal distribution. We may, therefore, consider that the variable “age” may be normally distributed in the population.

Quartile and Percentile: When a dataset is divided into four equal parts after arranging into ascending order, each part is called a quartile. It is expressed as Q1 (first quartile or 25th percentile), Q2 (second quartile or median or 50th percentile) and Q3 (third quartile or 75th percentile). On the other hand, when data is divided into 100 equal parts (after ordered array), each part is called a percentile.

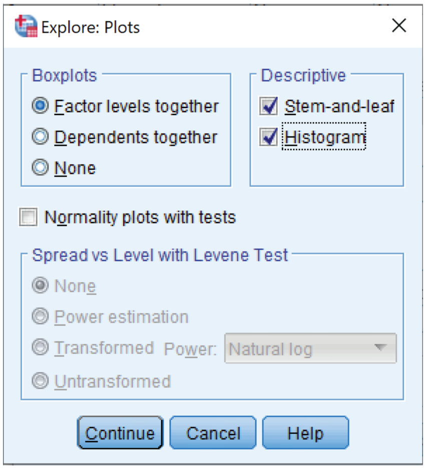
We can see in Table 4.4 that Percentile 25 (Q1), Percentile 50 (Q2) and Percentile 75 (Q3) for age are 21, 27 and 32 years, respectively. Q1 or the first quartile is 21 years, means that 25% of the study subjects’ age is less than or equal to 21 years. On the other hand, 30th percentile (P_{30}) is 22.3 years, which means that 30% of the study subjects’ age is less than or equal to 22.3 years. Hope, you can now interpret the P_{40}.

4.3 Central tendency and dispersion: Alternative method

If you want to generate all the descriptive statistics (central tendency and dispersion) and charts (such as histogram, stem and leaf, and box and plot charts) for the variable “age”, use the following commands.

> Analyze > Descriptive statistics > Explore > Select the variable “age” and push it into the "Dependent List” box > Plots > Select "Stem and leaf” and “Histogram" > Continue > OK (Fig 4.6)

Figure 4.6



Explore: Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
☑ Stem-and-leaf
☑ Histogram

□ Normality plots with tests

Spread vs Level with Levene Test
● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

4.3.1 Outputs

The outputs are shown in Table 4.5 and Figures 4.7 to 4.9. Figure 4.10 is the additional box and plot chart of diastolic BP.

Table 4.5 Descriptive statistics of age

			Statistic	Std. Error
AGE	Mean		26.5143	.51689
	95% Confidence Interval for Mean	Lower Bound	25.4953	
		Upper Bound	27.5333	
	5% Trimmed Mean		26.5608	
	Median		27.0000	
	Variance		56.107	
	Std. Deviation		7.49049	
	Minimum		6.00	
	Maximum		45.00	
	Range		39.00	
	Interquartile Range		11.0000	
	Skewness		-.092	.168
	Kurtosis		-.288	.334

Figure 4.7 Histogram of age

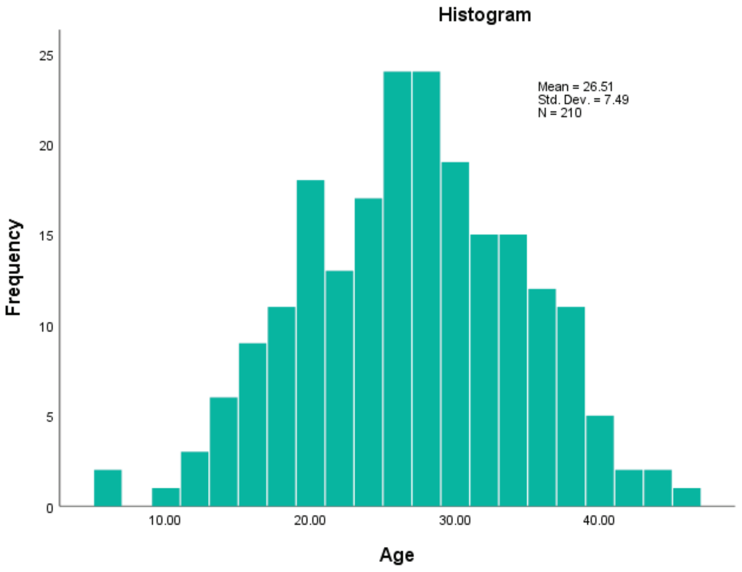


Figure 4.8 Stem and leaf chart of age

```
Age Stem-and-Leaf Plot
 Frequency    Stem & Leaf

    2.00        0 .  66
   10.00        1 .  0122344444
   24.00        1 .  555566666777788888889999
   44.00        2 .  00000000000000111112222222233333334444444444
   63.00        2 .  555555555666666666666666777777777788888888888888899999999999999
   34.00        3 .  0000111111112222222333333333444444
   26.00        3 .  55556666666677777788888999
    6.00        4 .  001133
    1.00        4 .  5

 Stem width:     10.00
 Each leaf:       1 case(s)
```

Figure 4.9 Box and plot chart of age

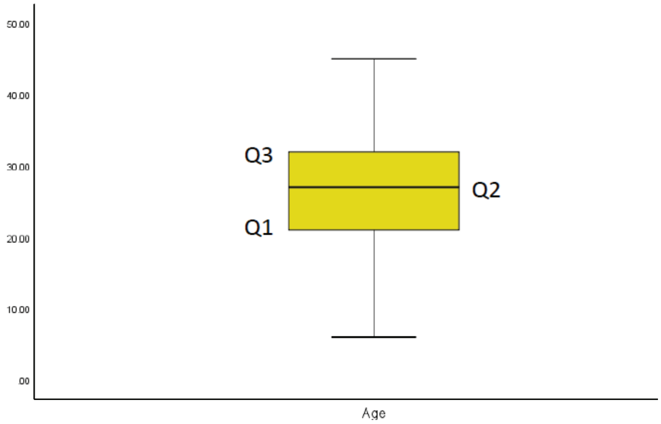
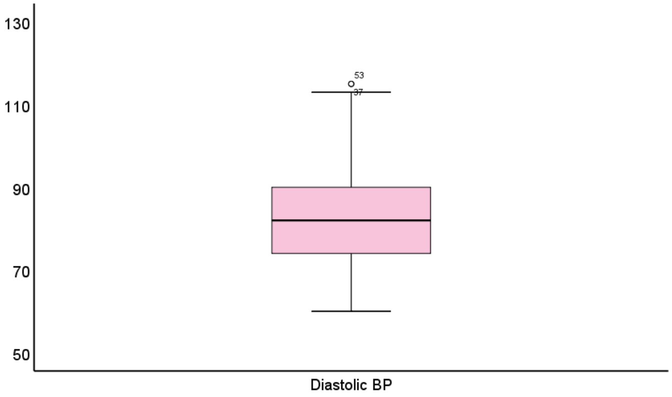


Figure 4.10 Box and plot chart of diastolic BP



4.3.2 Interpretation

Before we understand the graphs, let's review the descriptive statistics provided in Table 4.5. We can see that the SPSS has provided mean and *5% trimmed mean* of age. Five percent trimmed mean is the mean after discarding 5% of the upper and 5% of the lower values of age. The extent of the effect of outliers can be checked by comparing the mean with the 5% trimmed mean. If they are close together (as we see in Table 4.5; mean= 26.51 and 5% trimmed mean= 26.56), there is no significant influence of the outliers (or there are no outliers) on age in the dataset. If they are very different, it means that the outliers have a significant influence on the mean value, and suggests we need to check the outliers and extreme values in the dataset. Table 4.5 also shows the 95% Confidence Interval (CI) for the mean of "age", which is 25.49-27.53. The 95% CI for the mean indicates that we are 95% confident/sure that the mean age of the population lies between 25.49 and 27.53 years.

SPSS has provided several graphs (Figs 4.7 to 4.9), such as histogram, stem and leaf, and box and plot charts. The histogram gives us information about:

a) Distribution of the dataset (whether symmetrical or not);
b) Concentration of values (where most of the values are concentrated); and
c) Range of values.

Looking at the histogram (Fig 4.7), it seems that the data is more or less symmetrical. This indicates that age may be normally (approximately) distributed in the population.

Stem and leaf chart (Fig 4.8) provides information similar to a histogram but retains the actual information on data values. Looking at the stem and leaf chart, we can have an idea about the distribution of the dataset (whether symmetrical or not). Data displayed in Figure 4.8 shows that the data is more or less symmetrical. *Stem and leaf charts are suitable for small datasets.*

The box and plot chart (Fig 4.9) provides information about the distribution of a dataset. It also provides summary statistics of a variable, like Q1 (first quartile or 25th percentile), median (second quartile or Q2) and Q3 (third quartile or 75th percentile) as well as information about outliers/extreme values. The lower boundary of the box indicates the value for Q1, while the upper boundary indicates the value for Q3. The median is represented by the horizontal line within the box. The smallest and largest values are indicated by the horizontal lines of the whiskers.

In the box and plot chart, the presence of *outliers* is indicated by the ID number and circle, while the presence of *extreme values* is indicated by "*". Outliers are the values lying between 1.5 and <3 box length distance from the edge (upper or lower) of the box. On the other hand, the extreme values are 3 or more box length distance from the upper or lower edge of the box. Figure 4.9 shows that there is no outlier in the data for age. We have provided another box and plot chart, which is for the variable “diastolic BP” (Fig 4.10). Figure 4.10 shows that there are 2 outliers (ID no. 37 and 53) in the data of diastolic BP, but does not have any extreme value.

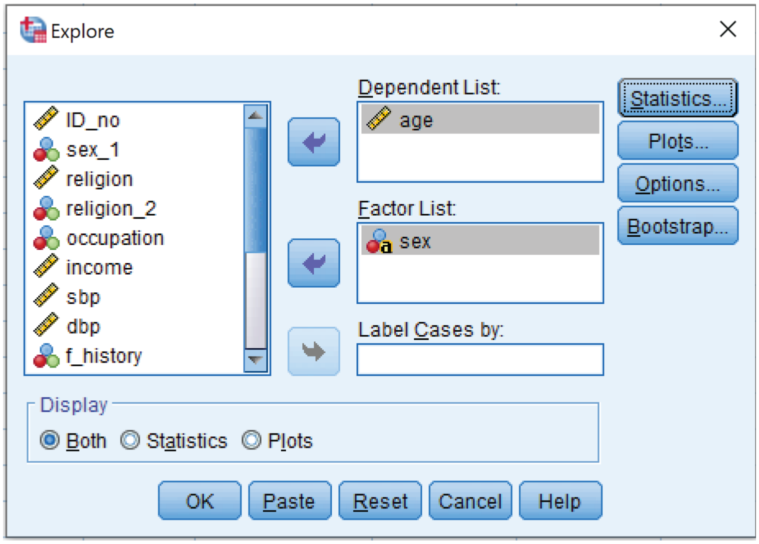
4.4 Descriptive statistics disaggregated by a categorical variable

You can get the descriptive statistics and other measures disaggregated by categorical variables. For example, if you want to calculate the measures of central tendency and dispersion of age by sex (i.e., by males and females separately), use the following commands. SPSS will produce the outputs separately for males and females (Table 4.6). *Note that there are other ways of doing this.*

Analyze > Descriptive statistics > Explore > Select “age” and push it into the "Dependent list" box > Select “sex” and push it into the "Factor list" box (Fig 4.11) > Plots > Deselect "Stem and leaf”, and select “Histogram" > Continue > OK

With these commands SPSS will provide the descriptive statistics by sex (Table 4.6) along with the graphs (only the descriptive statistics table is provided).

Figure 4.11



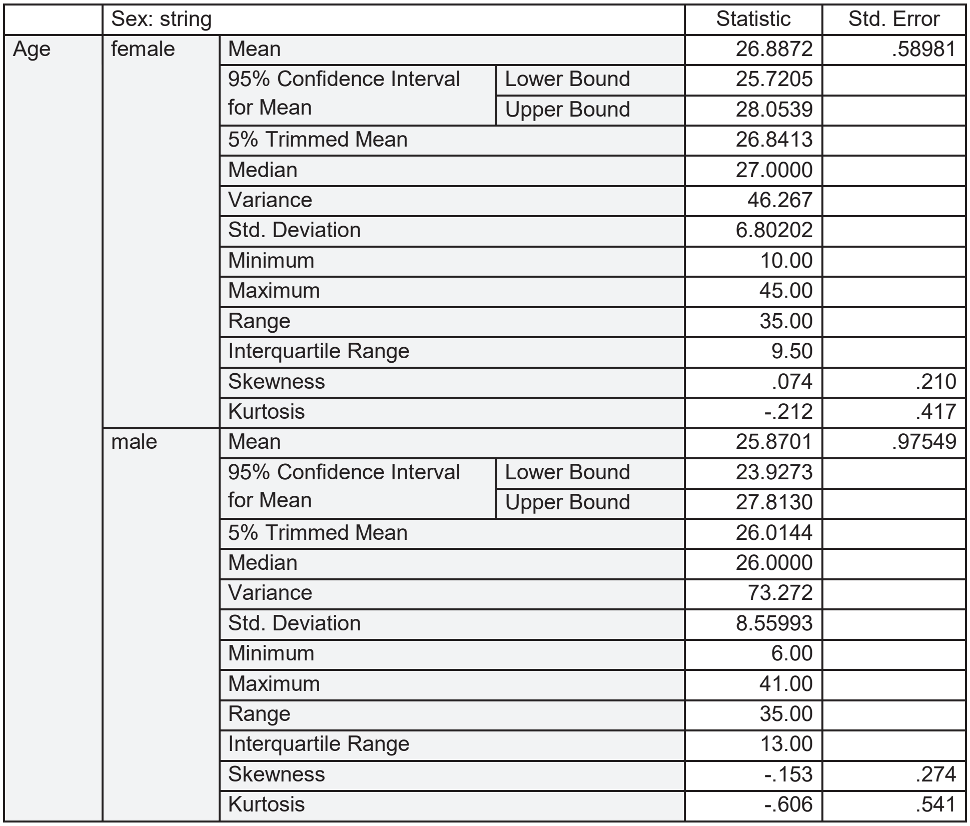
Dependent List: age

Factor List: sex

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Table 4.6 Descriptive statistics of age by sex



	Sex: string			Statistic	Std. Error
Age	female	Mean		26.8872	.58981
		95% Confidence Interval for Mean	Lower Bound	25.7205	
			Upper Bound	28.0539	
		5% Trimmed Mean		26.8413	
		Median		27.0000	
		Variance		46.267	
		Std. Deviation		6.80202	
		Minimum		10.00	
		Maximum		45.00	
		Range		35.00	
		Interquartile Range		9.50	
		Skewness		.074	.210
		Kurtosis		-.212	.417
	male	Mean		25.8701	.97549
		95% Confidence Interval for Mean	Lower Bound	23.9273	
			Upper Bound	27.8130	
		5% Trimmed Mean		26.0144	
		Median		26.0000	
		Variance		73.272	
		Std. Deviation		8.55993	
		Minimum		6.00	
		Maximum		41.00	
		Range		35.00	
		Interquartile Range		13.00	
		Skewness		-.153	.274
		Kurtosis		-.606	.541

4.5 Checking for outliers

Outliers and extreme values can be checked by looking at the box and plot chart, as discussed earlier. We can also check the presence of outliers using the following commands. For example, we want to check if there are any outliers present in the variable “age”.

> Analyze > Descriptive Statistics > Explore > Select the variable “age” and push it into the "Dependent list” box > Select “ID_no” (ID no.) and push it into the "Label cases by” box > Select “Statistics” under “Display” section > Statistics > Select “Outliers” > Continue > OK

The SPSS will provide 5 upper and 5 lower values of age with the corresponding ID (serial no.) numbers, as shown in Table 4.7.

Table 4.7 Upper 5 and lower 5 values of age with ID numbers

			Case Number	Id number	Value
Age	Highest	1	154	154	45
		2	117	117	43
		3	119	119	43
		4	2	2	41
		5	166	166	41
	Lowest	1	127	127	6
		2	98	98	6
		3	36	36	10
		4	49	49	11
		5	34	34	12[a]

a. Only a partial list of cases with the value 12 are shown in the table of lower extremes.

5

Checking Data for Normality

It is important to know the nature of distribution of a continuous random variable before using statistical methods for hypothesis testing. To use the parametric methods for testing hypotheses (e.g., t-test, ANOVA), one of the important assumptions is that the data of the dependent variable are normally distributed. It is, therefore, necessary to check whether the data has been derived from a normally distributed population or not, before we use the parametric methods. Use the data file <**Data_3.sav**> for practice.

5.1 Assessing normality of data

This is an important assumption for using a parametric test. Whether the data has come from a normally distributed population or not, can be checked in several different ways. The commonly used methods are:

a) Graphs, such as histogram and Q-Q chart;
b) Descriptive statistics, using skewness and kurtosis; and
c) Formal statistical tests, such as 1-sample Kolmogorov Smirnov (K-S) test and Shapiro Wilk test.

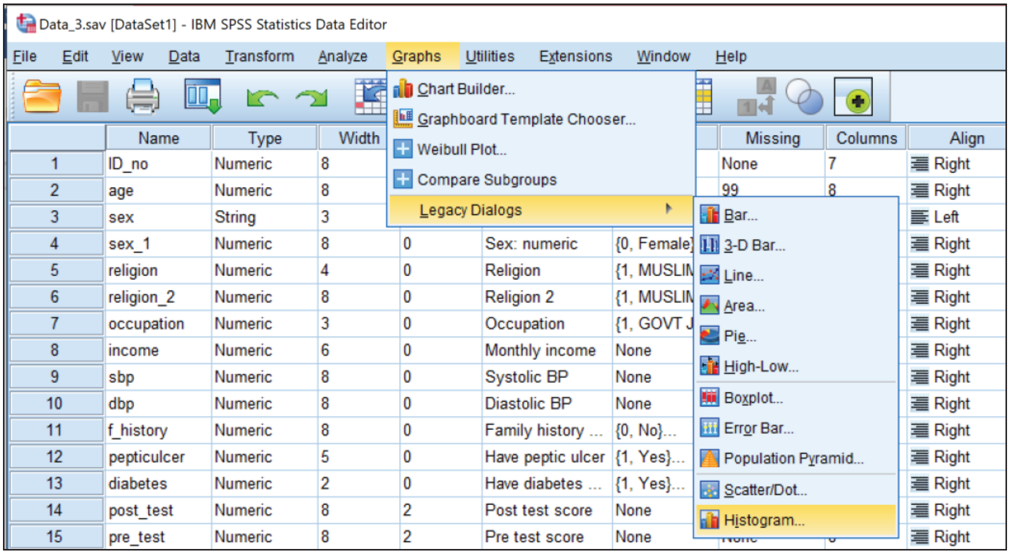
Let us see how to generate the histogram and Q-Q chart, and do the formal statistical tests (K-S test and Shapiro Wilk test).

Suppose we want to know whether the variable “systolic BP (SPSS variable name: sbp)” is normally distributed in the population or not. We shall first construct the histogram and Q-Q chart. To construct a histogram for systolic BP, use the following commands.

Graphs > Legacy dialogs > Histogram > Select the variable “sbp” and push it into the “Variable” box > Select “Display normal curve” clicking at the box below > OK (Fig 5.1)

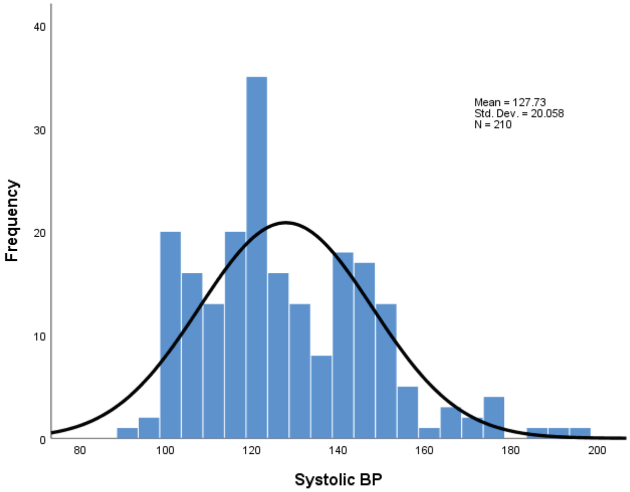
The SPSS will produce the histogram of systolic BP, as shown in Figure 5.2.

Figure 5.1



	Name	Type	Width				Missing	Columns	Align
1	ID_no	Numeric	8				None	7	Right
2	age	Numeric	8				99	8	Right
3	sex	String	3						Left
4	sex_1	Numeric	8	0	Sex: numeric	{0, Female}			Right
5	religion	Numeric	4	0	Religion	{1, MUSLIN			Right
6	religion_2	Numeric	8	0	Religion 2	{1, MUSLIN			Right
7	occupation	Numeric	3	0	Occupation	{1, GOVT J			Right
8	income	Numeric	6	0	Monthly income	None			Right
9	sbp	Numeric	8	0	Systolic BP	None			Right
10	dbp	Numeric	8	0	Diastolic BP	None			Right
11	f_history	Numeric	8	0	Family history ...	{0, No}...			Right
12	pepticulcer	Numeric	5	0	Have peptic ulcer	{1, Yes}...			Right
13	diabetes	Numeric	2	0	Have diabetes ...	{1, Yes}...			Right
14	post_test	Numeric	8	2	Post test score	None			Right
15	pre_test	Numeric	8	2	Pre test score	None			Right

Figure 5.2 Histogram of systolic BP

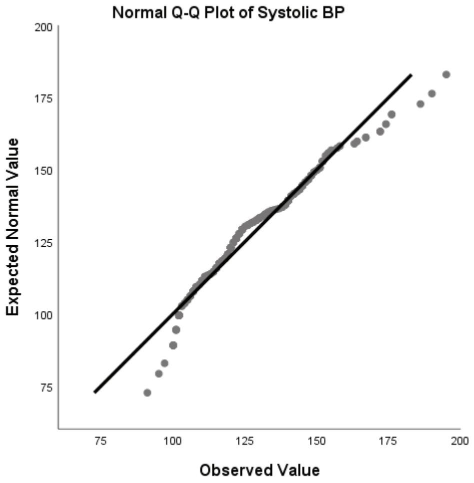


To generate the Q-Q plot for systolic BP, use the following commands.

Analyze > Descriptive statistics > Q-Q Plots > Select the variable "sbp" and push it into the "Variables" box > For Test Distribution select "Normal" (usually remains as default) > OK

SPSS will produce a Q-Q plot for systolic BP, as shown in Figure 5.3.

Figure 5.3 Q-Q plot of Systolic BP



To do the formal statistical tests (K-S test and Shapiro Wilk test) to understand the normality of the data, use the following commands.

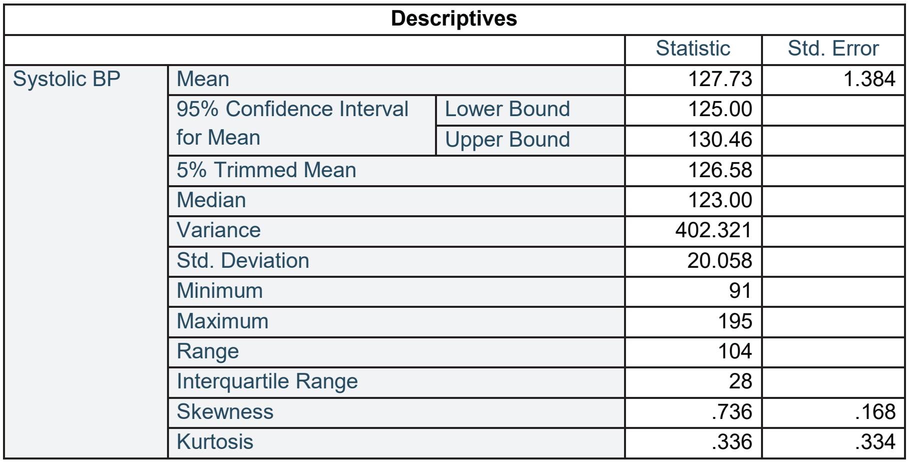
Analyze > Descriptive Statistics > Explore > Select the variable “sbp” and push it into the "Dependent List" box > Plots > Deselect "Stem-and-leaf" and select “Histogram” > Select “Normality plots with test” > Continue > OK

Note that these commands will also produce the histogram and Q-Q plot. You may not need to develop the histogram and Q-Q plot separately as mentioned earlier.

5.1.1 Outputs

You will be presented with the following tables (Table 5.1 and 5.2) along with the histogram, Q-Q plot and box and plot chart. The histogram, Q-Q plot and box and plot chart, generated by the commands, have been omitted to avoid repetition.

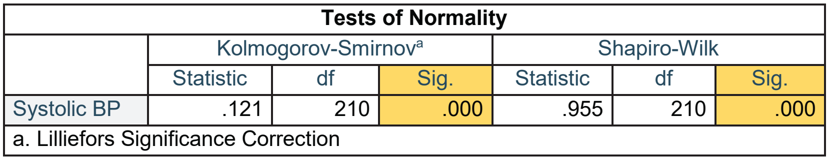
Table 5.1 Descriptive statistics of Systolic BP



Descriptives

			Statistic	Std. Error
Systolic BP	Mean		127.73	1.384
	95% Confidence Interval for Mean	Lower Bound	125.00	
		Upper Bound	130.46	
	5% Trimmed Mean		126.58	
	Median		123.00	
	Variance		402.321	
	Std. Deviation		20.058	
	Minimum		91	
	Maximum		195	
	Range		104	
	Interquartile Range		28	
	Skewness		.736	.168
	Kurtosis		.336	.334

Table 5.2 Statistical tests for normality of data (of Systolic BP)



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Systolic BP	.121	210	.000	.955	210	.000

a. Lilliefors Significance Correction

5.1.2 Interpretation

SPSS has generated the histogram and Q-Q plot (Fig 5.2 and 5.3) for "systolic BP" and Tables 5.1 and 5.2. While performing the specific statistical tests (K-S test and Shapiro Wilk test) to check the normality of a dataset, SPSS automatically provides the descriptive statistics of the variable (Table 5.1). We have already discussed the measures of skewness and kurtosis to assess the normality of a dataset earlier (chapter 4).

Histogram (Fig 5.2) provides an impression about the distribution of the dataset (whether the distribution is symmetrical or not). If we look at the histogram of systolic BP, it seems that the data is slightly skewed to the right (i.e., distribution is not symmetrical).

The Q-Q plot (Fig 5.3) also provides information on whether data have come from a normally distributed population or not. The Q-Q plot compares the distribution of data with the standardized theoretical distribution from a specified family of distribution (in this case from normal distribution). *If data are normally distributed, all the points (dots) lie on the straight line*. Note that our interest is in the central portion of the line. Deviation from the central portion of the line means non-normality. Deviations at the ends of the plot indicate the existence of outliers. We can see (in Fig 5.3) that there is a slight deviation of the dots at the central portion as well as at the two ends. This may indicate that the data may not have come from a normally distributed population.

The specific tests (objective tests) to assess if the data have come from a normally distributed population are the *K-S (Kolmogorov-Smirnov) test and Shapiro Wilk test*. The results of these two tests are provided in Table 5.2.

Look at the Sig (significance) column of Table 5.2. Here, Sig (indicates the p-value) is 0.000 for both the tests. A p-value of <0.05 indicates that the data *have not* come from a normally distributed population. In our example, the p-value is 0.000 for both the tests, which is <0.05. This means that the data of systolic BP have not come from a normally distributed population. Here the *null hypothesis* is "data have come from a normally distributed population". The alternative hypothesis is "data have not come from a normally distributed population". We will reject the null hypothesis, since the p-values of the tests are <0.05.

Note that the K-S test is very sensitive to sample size. The K-S test may be significant for slight deviations of a large sample data (n>100). Similarly, the likelihood of getting a p-value <0.05 for a small sample (n<20, for example) is low. Therefore, the rules of thumb for normality checking are:

1) Sample size <30: Assume non-normal;
2) Moderate sample size (30-100): If the formal test is significant (p<0.05), consider non-normal distribution, otherwise check by other methods, e.g., histogram, and Q-Q plot; and
3) Large sample size (n>100): If the formal test is not significant (p>0.05), accept normality, otherwise check by other methods.

For practical purposes, just look at the histogram. If it seems that the distribution is approximately symmetrical, consider that the data have come from a normally distributed population.

6

Data Management

While analyzing data, you may require to make class intervals, classify a group of people with a specific characteristic using a cutoff value (e.g., you may want to classify people who have hypertension using a cutoff value of either systolic or diastolic BP), and recode data for other purposes. In this chapter, we will discuss data manipulations that are commonly needed during data analysis, such as:

- Recoding of data
- Making class intervals
- Combine data to form an additional variable
- Data transformation
- Calculation of total score
- Extraction of time
- Selection of a subgroup for data analysis
- Split file for data analysis
- Making variable sets

Use the data file <**Data_3.sav**> for practice.

6.1 Recoding of data

In the dataset, the variable "sex" is coded as "m" and "f". You want to replace the existing code "m" by 1 and "f" by 2. There are two options for recoding of data:

a) Recoding into same variable; and
b) Recoding into different variable.

My suggestion is to use the "recoding into different variable" option all the time. This will keep the original data of the variable intact.

6.1.1 Recoding into same variable

Note that if you recode data into same variable, the original data/coding will be lost. Use the following commands to recode data into the same variable.

> Transform > Recode into same variables > Select "sex" and push it into the "Variables" box > Click on "Old and new values" > Select "Value" (usually default) under "Old value" section > Type **m** in the box below > type 1 in the "Value" area under "New value" section > Click "Add" > Type **f** in the "Value" area under "Old value" section > type 2 in the "Value" area under "New value" section > Click "Add" > Continue > OK (Fig 6.1 to 6.3)

Figure 6.1

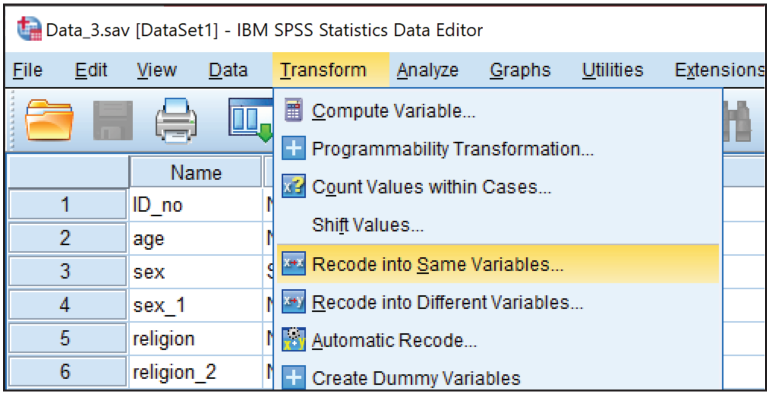


Figure 6.2

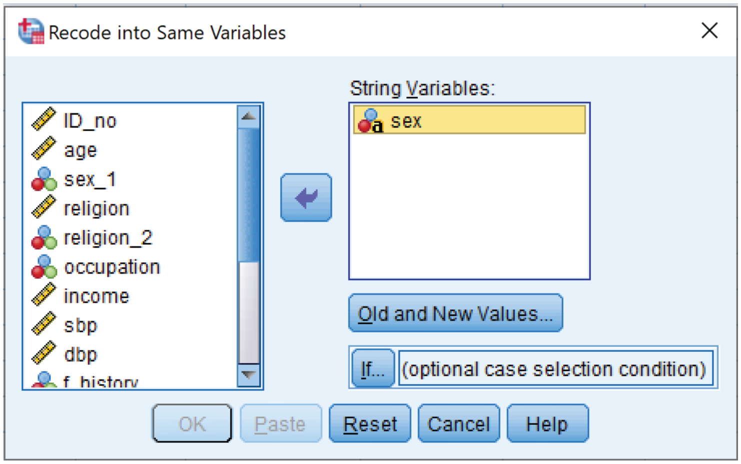
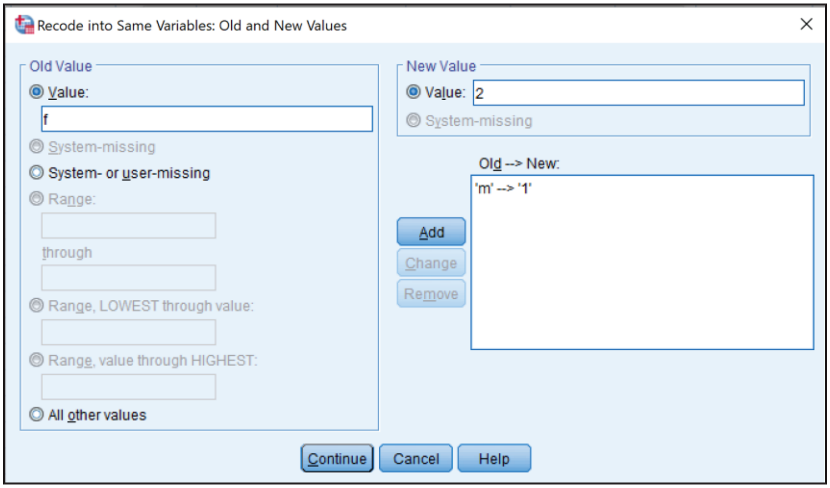


Figure 6.3



Recode into Same Variables: Old and New Values

Old Value

● Value: f

○ System-missing

○ System- or user-missing

○ Range:

through

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

○ All other values

New Value

● Value: 2

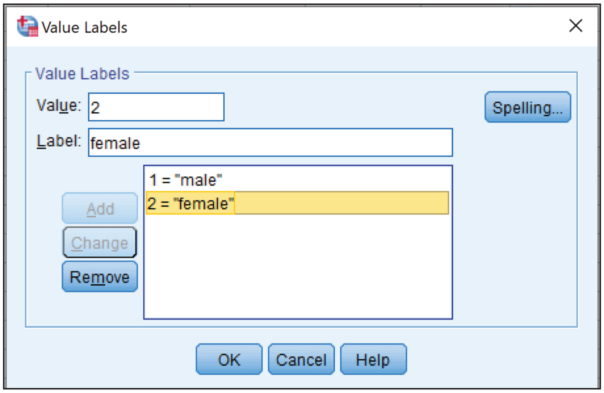
○ System-missing

Old --> New:

'm' --> '1'

Check the data file in the "data view" option. You will notice that all the "m" has been replaced by 1 and "f" by 2. Now, go to the "variable view" of the data file to replace the codes. Click in the "Values" box against the variable "sex". Replace the codes as 1 is "Male" and 2 is "Female" (Fig 6.4).

Figure 6.4



Value Labels

Value Labels

Value: 2

Label: female

1 = "male"
2 = "female"

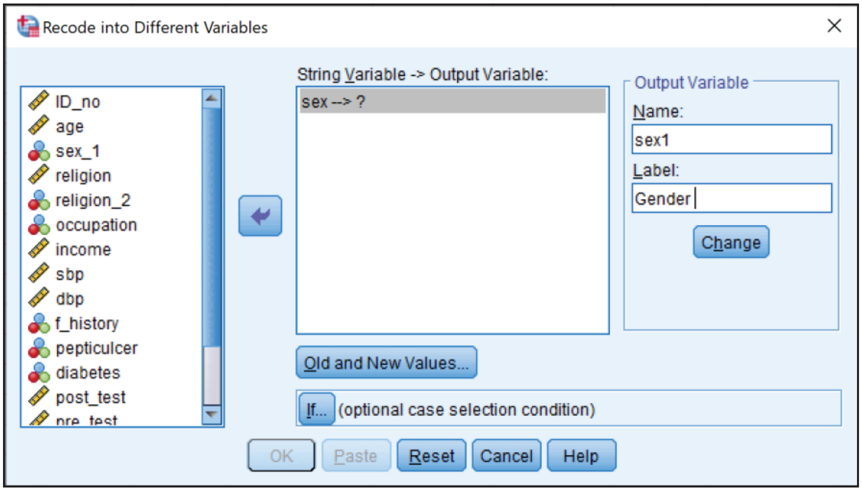
6.1.2 Recoding into different variable

This option of recoding requires formation of a new variable. The original variable and data will remain intact. To do this, use the following commands.

Transform > Recode into different variables > Select "sex" and push it into "Input Variable –Output Variable" box > Type "sex1" in the "Name" box and type "Gender" in the “Label” box under “Output variable” section > Click “Change” > Click “Old and new values” > Type **m** in the “Value” box under the “Old value” section > Type 1 in the “Value” box under the “New value” section > Click “Add” > Type **f** in the “Value” box under the “Old value” section > Type 2 in the “Value” box under the “New value” section > Click “Add” > Continue > OK (Fig 6.5 and 6.6)

With these commands, the new variable generated is “**sex1**” (do not allow any space between “sex” and 1, while typing the variable name in the Name box). Follow the rules of writing variable names as mentioned in chapter 2 (section 2.1.1).

Figure 6.5

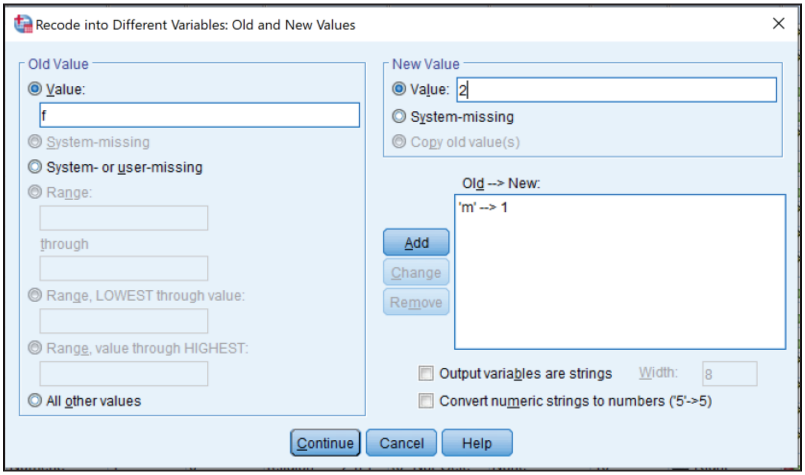


Recode into Different Variables

String Variable -> Output Variable:
sex --> ?

Output Variable
Name: sex1
Label: Gender

Figure 6.6



Recode into Different Variables: Old and New Values

Old Value
● Value: f
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values

New Value
● Value: 2
○ System-missing
○ Copy old value(s)

Old --> New:
'm' --> 1

□ Output variables are strings Width: 8
□ Convert numeric strings to numbers ('5'->5)

Click on the “variable view” option of the data file. You will notice that SPSS has generated a new variable “sex1” (the last variable both in the variable view and data view options). Like before (sections 2.1.1 and 6.1.1), in the variable view, define the value labels of the new variable “sex1” as 1 is “male” and 2 is “female” (Fig 6.4).

6.2 Converting string variables into numeric variables

The SPSS data file may have string as well as numeric variables. It is always preferable to generate numeric variables rather than string variables. Numeric variables are easier to manipulate and can be used in various statistical analyses. Some analysis in SPSS does not allow string variables, e.g., one-way ANOVA does not allow a string variable for its factor option. A string variable may be coded or have a direct response (e.g., name of district or province). The codes of a string variable may be a character (e.g., m= male; f= female) or a number (e.g., 1= male; 2= female). Even though the codes are with numbers, they are actually the characters. We can easily convert a string variable into a numeric variable by SPSS.

In our dataset, “sex” is a string variable coded as “f= female” and “m= male”. Let us convert the variable “sex” into a numeric variable “sex1”. Use the following commands to do this.

Transform > Automatic recode > Select “sex” and push it into “Variable-new name” box > Write “sex1” in “New Name” box > Click on “Add new name” > OK

This will generate a new numeric variable “sex1” with codes “1= Female” and “2= Male”. You can find this variable at the bottom of the variable view of the dataset.

6.3 Making class intervals

Central tendency (such as mean, median) and dispersion (such as SD) of quantitative data provide meaningful information. Further useful summarization may be achieved by grouping the data into class intervals or categories. For instance, you want to categorize the variable “age” into the following categories/class intervals:

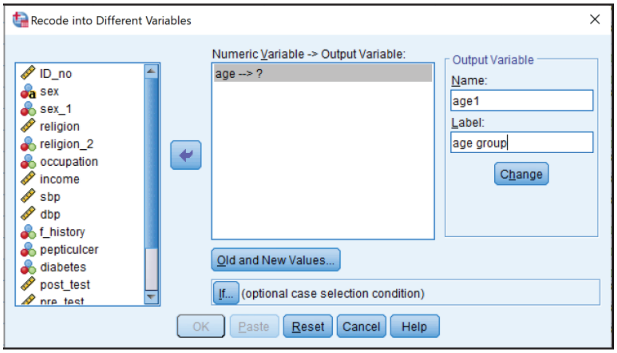
- ≤20 years (to be coded as 1),
- 21-30 years (to be coded as 2),
- 31-40 years (to be coded as 3), and
- >40 years (to be coded as 4).

We shall use the option “Recode into Different Variable” for this exercise. *We would suggest the users to select the option “Recode into Different Variables” all the times.* If you use the option “Recode into Same Variables”, you will lose the original data that cannot be recovered once the data file is saved. As stated before, we need to generate a new variable for this option. Suppose the new variable we want to generate is “age1”. To recode the variable “age” into “age1”, use the following commands.

> Transform > Recode into Different Variables > Select “age” and push it into the “Input Variable–Output Variable” box > Type “age1” in the “Name” box and type “age group” in the “Label” box under “Output Variable” section > Click “Change” > Click on “Old and New Values” > Select “System-missing” under “Old value” section > Select “System-missing” under “New Value” section > Click “Add” > Select “Range, LOWEST through value” and type “20” in the box below > Select “Value” under “New Value” section and type “1” > Add > Now select “Range” under “Old Value” section > Type “21” in the upper box and “30” in the lower box > Select “Value” under “New Value” section and type “2” > Add > Again, type “31” in the upper box and “40” in the lower box > Select “Value” under “New Value” section and type “3” > Add > Select “All other values” under “Old Value” section > Select “Value” under “New Value” section and type “4” > Add > Continue > OK (Fig 6.7 and 6.8)

These commands will generate a new variable “age1” with four categories of the variable “age” as 1, 2, 3, and 4. Go to the data file in the “data view” option and then to the “variable view” option. You will notice that SPSS has generated a new variable “age1” (the last variable both in the data view and variable view options) with the values 1 to 4. Like before, in the “variable view” option, define the value labels of the variable “age1” as 1 is “≤ 20 years”, 2 is “21-30 years”, 3 is “31-40 years” and 4 is “>40 years”. Now, make a frequency distribution of the variable “age1”. SPSS will provide the following table (Table 6.1).

Figure 6.7



Recode into Different Variables

Numeric Variable -> Output Variable:

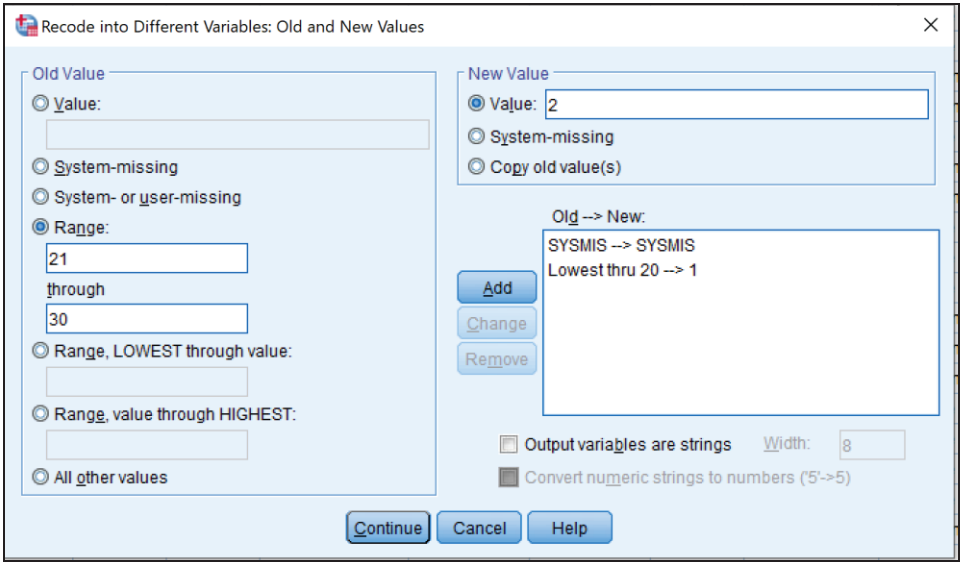
age --> ?

Output Variable

Name: age1

Label: age group

Figure 6.8



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

○ System- or user-missing

● Range:

21

through

30

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

○ All other values

New Value

● Value: 2

○ System-missing

○ Copy old value(s)

Old --> New:

SYSMIS --> SYSMIS
Lowest thru 20 --> 1

□ Output variables are strings Width: 8

□ Convert numeric strings to numbers ('5'->5)

Table 6.1 Frequency distribution of age group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< = 20 years	50	23.8	23.8	23.8
	21-30 years	97	46.2	46.2	70.0
	31-40 years	58	27.6	27.6	97.6
	> 40 years	5	2.4	2.4	100.0
	Total	210	100.0	100.0	

Using the transform command, you can also classify individuals who have hypertension and who do not have hypertension (as an example). To do this, you shall have to use a cutoff value to define hypertension. For example, we have collected data on diastolic BP (SPSS variable name is "dbp"). We want to classify those as "hypertensive" if the diastolic BP is >90 mmHg. Now, recode the variable diastolic BP into a new variable (say, d_hyper) using "Recode into Different Variables" option as ≤ 90 (normal BP) and > 90 (as hypertensive). Hopefully, you can do it now. If you cannot, use the following commands.

Transform > Recode into Different Variables > Select "dbp" and push it into the "Input Variable –Output Variable" box > Type "d_hyper" in the "Name" box and type "diastolic hypertension" in the "Label" box under "Output Variable" section > Click "Change" > Click on "Old and New Values" > Select "System-missing" under "Old value" section > Select "System-missing" under "New Value" section > Add > Select "Range, LOWEST through value" under "Old value" section and type "90" in the box below > Select “Value” under “New Value” section and type “1” > Add > Select “All other values” under “Old value” section > Select “Value” under “New Value” section and type “2” > Add > Continue > OK

This will create a new variable “d_hyper” with code numbers 1 and 2 (the last variable both in the variable and data view options). Code 1 indicates the persons without hypertension (diastolic BP ≤ 90) and code 2 indicates the persons with hypertension (diastolic BP >90). As discussed previously, in the “variable view” option, define the value labels of the new variable “d_hyper” as 1 is “Do not have hypertension” and 2 is “Have hypertension”. Make a frequency distribution of the new variable “d_hyper” to find the proportion of subjects with diastolic hypertension.

6.4 Combine data into a new variable

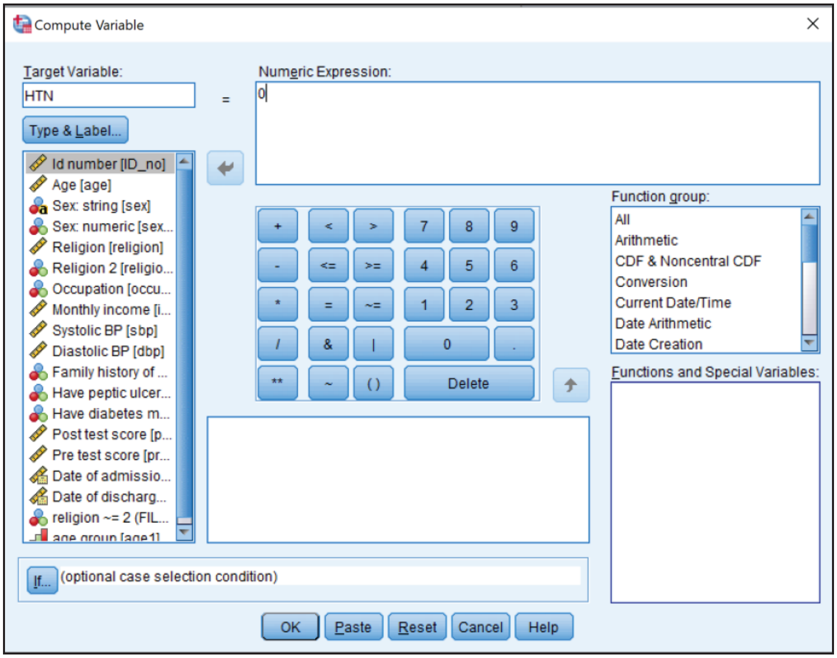
Sometimes, the cutoff point of a measurement (e.g., hemoglobin, blood pressure) for defining a condition (e.g., anemia, hypertension) may vary according to gender or other characteristics. In such a situation, a single cutoff point for defining a condition may not be appropriate.

For example, we have collected data on diastolic BP (SPSS variable name is “dbp”) for male and female patients. We have defined hypertension as diastolic BP >85 mmHg if it is a female, and diastolic BP >90 mmHg if it is a male. Now, how to classify those who have hypertension based on gender?

To do this, first, we shall generate a new variable, say “HTN” for which all the values will be 0 (zero). Use the following commands to do this.

Transform > Compute Variable > Type “HTN” in “Target Variable” box > Click in the box under “Numeric Expression” > Write “0” (zero) using the number pad or keyboard > OK (Fig 6.9)

Figure 6.9



Compute Variable

Target Variable: HTN

Numeric Expression: 0

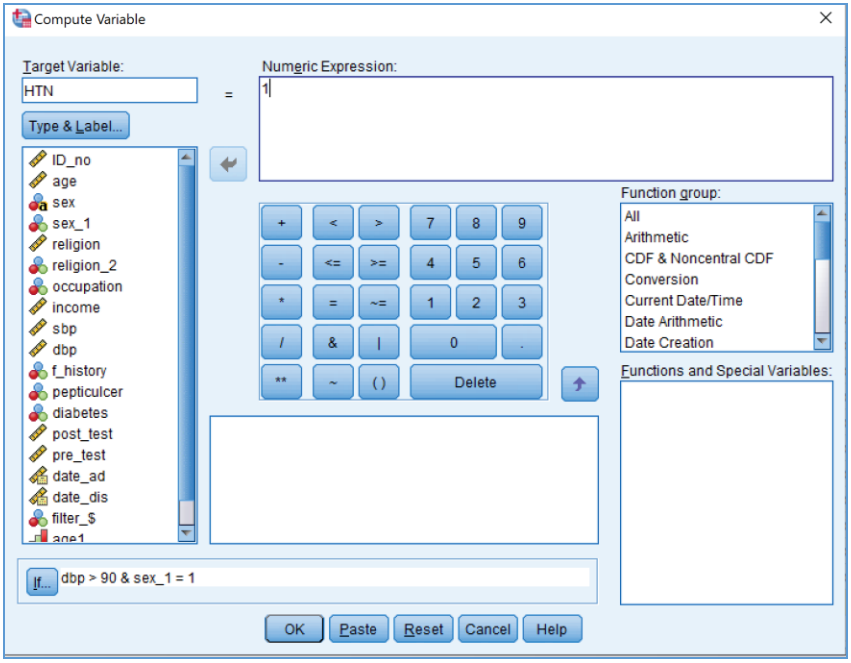
This will generate the new variable "HTN" with all the values 0 (you can check it in the "data view" option; the last variable). Now use the following commands.

Transform > Compute Variable > Click in the box under "Numeric Expression" > Delete 0 > Click 1 on the "number pad" (or type 1 using the keyboard) > Click "If (optional case selection condition)" > Select "Include if case satisfies condition" > Select "dbp" and push it into the box > Click "greater than sign (>)" then write "90" using the "number pad" (always use the number pad) > Click "&" on the "number pad" > Select "sex_1" and push it into the box > Click on "=" and then "1" (note: 1 is the code no. for male) > Continue > OK > SPSS will provide the message "Change existing variable?" > Click on "OK" (Fig 6.10 and 6.11)

Again,

Transform > Compute Variable > Click "If (optional case selection condition)" > Delete "90" and write "85" (for dbp) and delete "1" and write "0" (for sex_1, since 0 is the code for female) > Continue > OK > SPSS will give you the message "Change existing variable?" > Click "OK"

Figure 6.10



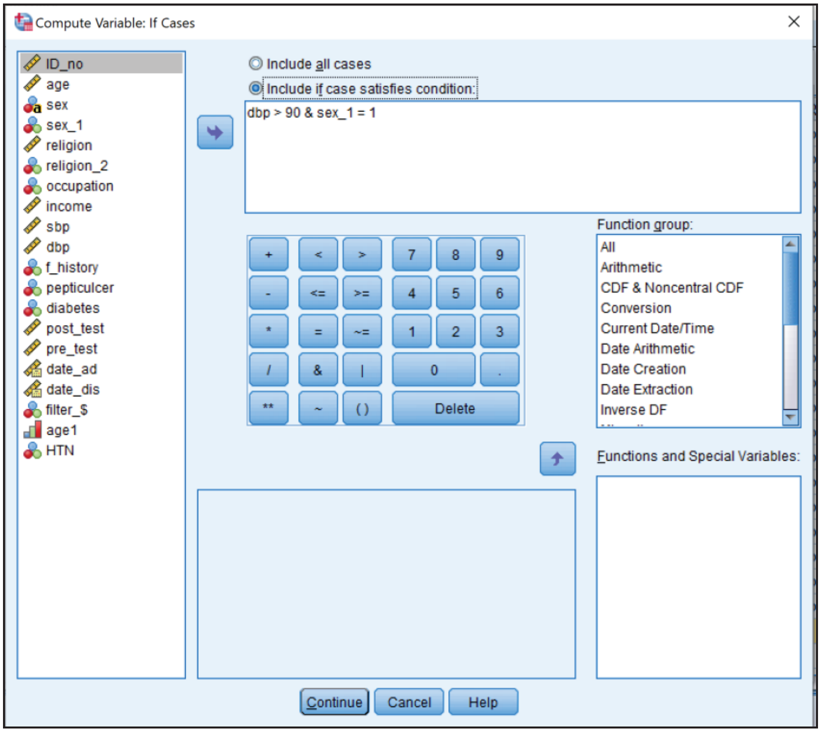
Compute Variable

Target Variable: HTN

Numeric Expression: 1

If... dbp > 90 & sex_1 = 1

Figure 6.11



Compute Variable: If Cases

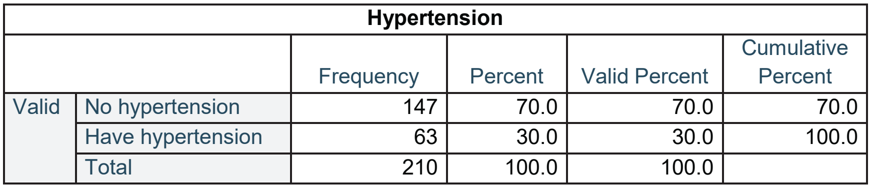
○ Include all cases

● Include if case satisfies condition:

dbp > 90 & sex_1 = 1

Go to the “data view” option of the data file. You will notice that the new variable “HTN” (the last variable both in the “data view” and “variable view” options) has values either “0” or “1”. “0” indicates “no hypertension”, while “1” indicates “have hypertension”. Like before, go to the “variable view” option and define the value labels of the variable “HTN” as “0” is “No hypertension” and “1” is “Have hypertension”. Now make a frequency of the variable “HTN” to determine the proportion of subjects with hypertension (Table 6.2).

Table 6.2 Frequency distribution of Hypertension



Hypertension

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No hypertension	147	70.0	70.0	70.0
	Have hypertension	63	30.0	30.0	100.0
	Total	210	100.0	100.0	

6.5 Data transformation

Frequently, the data that we collect for our studies are not normally distributed. Since parametric methods (in general) for testing hypotheses are more conclusive than the non-parametric methods, data transformations are occasionally needed to make the distribution normal and to meet the assumptions for a parametric test. Depending on the shape of the data distribution, there are several options for data transformation. The following table (Table 6.3) shows some of the options for data transformation.

Table 6.3 Data transformation options

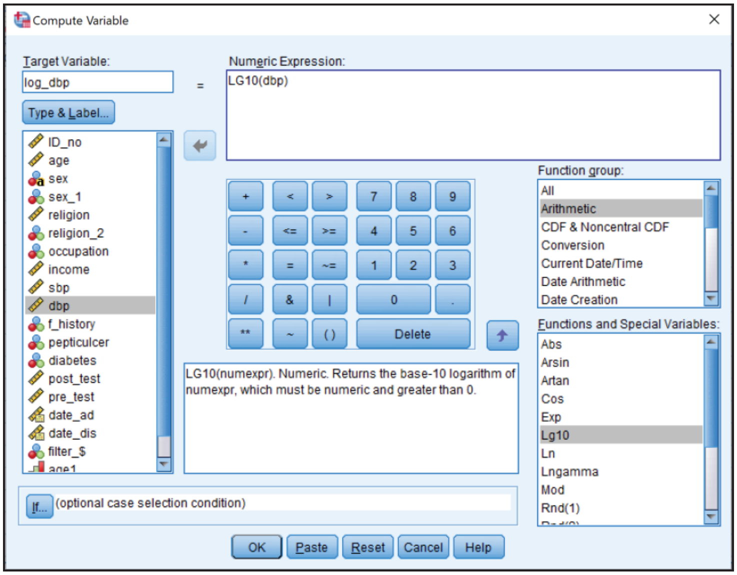
Method	Good for	Bad for
Log	Right skewed data	Zero values and negative values
Square root	Right skewed data	Negative values
Square	Left skewed data	Negative values
Reciprocal	Making small values bigger and big values smaller	Zero values and negative values

A commonly used method of data transformation is Log transformation. Let us see how to produce the Log transformation of data. Suppose you want to transform diastolic BP (variable name is “dbp”) into Log of diastolic BP. Use the following commands.

Transform > Compute Variable > Type “log_dbp” in the box under “Target Variable” > Click on “Arithmetic” in the “Function Group” box > In “Functions and Special Variables” box select “Lg10” > Click on the “up arrow” (left side of the box). You will see LG10(?) appears in the “Numeric Expression” box > Select “dbp” in the “Type and Label” box > Push it into the “Numeric Expression” box > OK (Fig 6.12)

This will generate a variable “log_dbp”, with the values “log of diastolic BP” (the last variable). Similarly, you can transform your data into square root using the option “sqrt” in the “Functions and special variables” box.

Figure 6.12



Target Variable: log_dbp

Numeric Expression: LG10(dbp)

Function group: Arithmetic

Functions and Special Variables: Lg10

LG10(numexpr). Numeric. Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.

6.6 Calculation of total score

Assume that you have conducted a study to assess the knowledge of secondary school children on how HIV is transmitted. To assess their knowledge, you have set the following questions (data file: **HIV.sav**).

HIV is transmitted through:

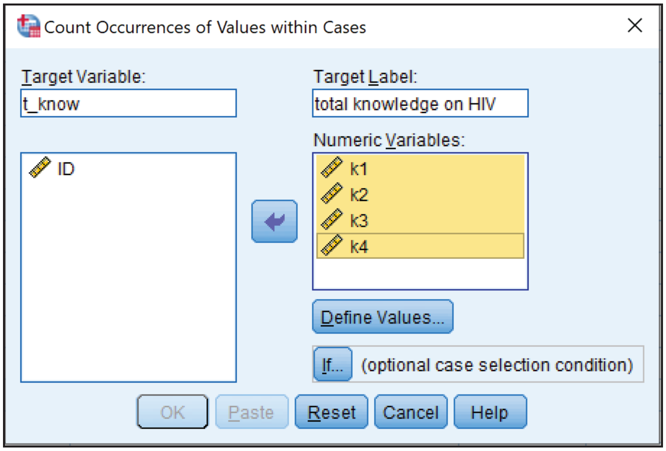
1. Sexual contact (variable name: **k1**) 1. Yes 2. No
2. Transfusion of unscreened blood (variable name: **k2**) 1. Yes 2. No
3. Sharing of injection needle (variable name: **k3**) 1. Yes 2. No
4. Accidental needle stick injury (variable name: **k4**) 1. Yes 2. No

Note: All the correct answers are coded as 1.

To calculate the total knowledge score, use the following commands.

Transform > Count values within cases > Write "t_know" in the box under "Target variable" > Write "total knowledge on HIV" in the box under "Target level" > Select "k1, k2, k3 and k4" and push them into the "Variables" box > Click "Define values" > Select "Value" and write "1" (since 1 is the correct answer) in the box below > Click "Add" > Continue > OK (Fig 6.13 and 6.14)

Figure 6.13



Count Occurrences of Values within Cases

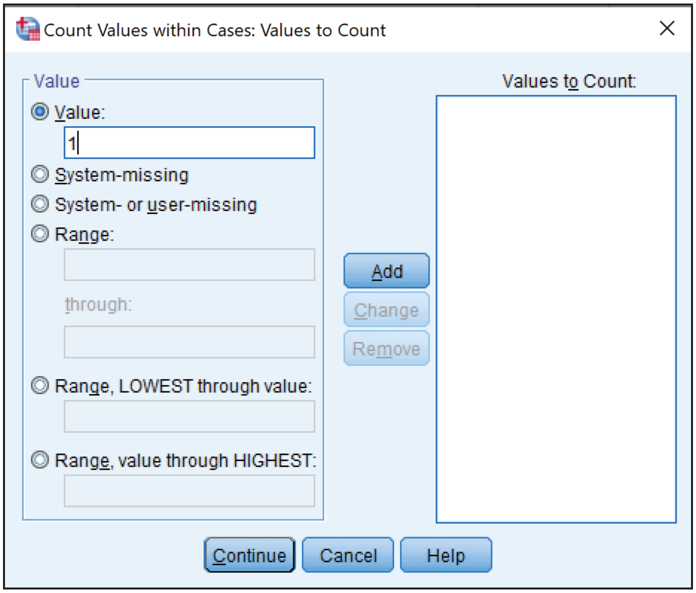
Target Variable: t_know

Target Label: total knowledge on HIV

ID

Numeric Variables:
- k1
- k2
- k3
- k4

Figure 6.14



Count Values within Cases: Values to Count

Value

● Value: 1

○ System-missing

○ System- or user-missing

○ Range:

through:

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

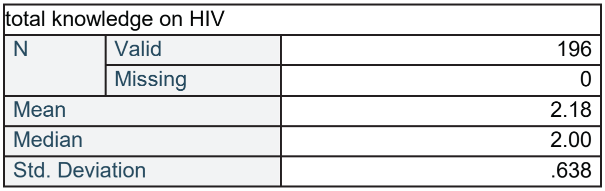
Values to Count:

SPSS will generate a new variable “t_know (total knowledge on HIV)” (look at the “variable view”). This variable has the total score of knowledge of the students. Now, you can obtain the descriptive statistics and frequency of the variable “t_know” by using the following commands.

Analyze > Descriptive statistics > Frequencies > Select “t_know” and push it into the “Variables” box > Statistics > Select “Mean, Median and Std. deviation” > Continue > OK

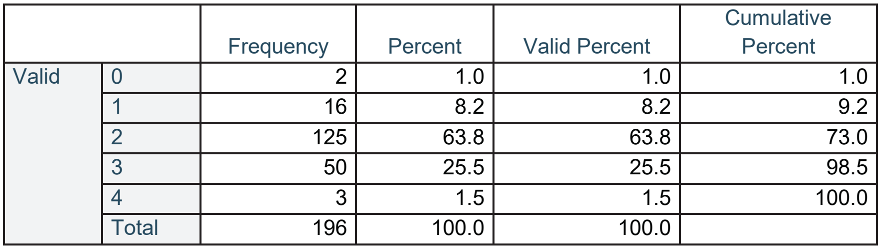
You will see the tables showing the descriptive statistics (mean, median) (Table 6.4) and frequency distribution of total knowledge (Table 6.5) of the students. Table 6.4 shows that the mean of the total knowledge is 2.18 (SD 0.63) and the median is 2.0. Table 6.5 shows that there are 2 (1%) students who do not have any knowledge on HIV transmission (since the score is 0, i.e., could not answer any question correctly). One hundred and twenty-five (63.8%) students know two ways of HIV transmission, while only 1.5% of the students know all the ways of HIV transmission. You can also classify the students as having “Good” or “Poor” knowledge using a cutoff value based on the total score.

Table 6.4 Descriptive Statistics of total knowledge



total knowledge on HIV		
N	Valid	196
	Missing	0
Mean		2.18
Median		2.00
Std. Deviation		.638

Table 6.5 Frequency distribution of total knowledge on HIV



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	1.0	1.0	1.0
	1	16	8.2	8.2	9.2
	2	125	63.8	63.8	73.0
	3	50	25.5	25.5	98.5
	4	3	1.5	1.5	100.0
	Total	196	100.0	100.0	

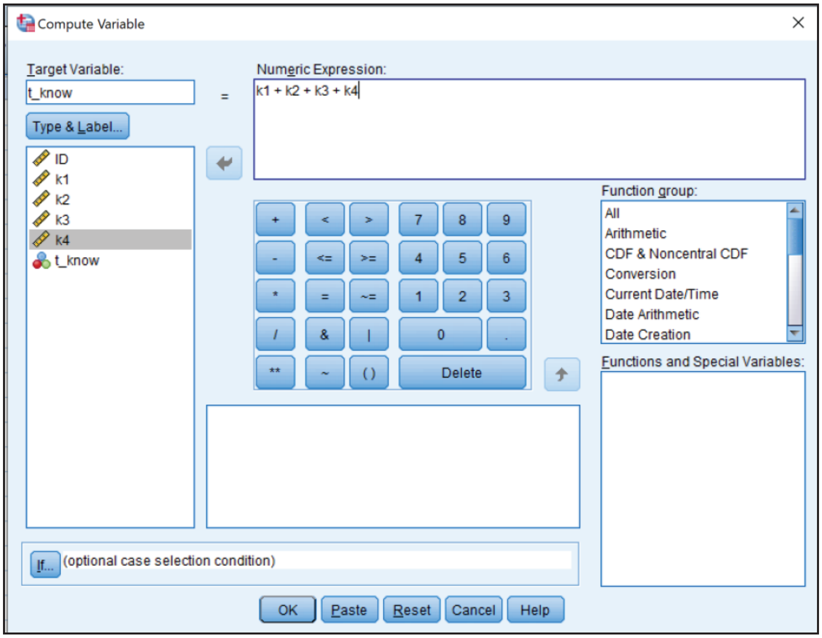
There is an alternative way of getting the total score. In that case, *the correct answers have to be coded as “1”, while the incorrect answers must be coded as “0” (zero)*. The commands are as follows.

Transform > Compute variable > Write “t_know” in the “Target variable” box > Select “k1” under ‘Type and label” and push it into the “Numeric expression” box >

From the key pad click “+” > Select “k2” and push it into the “Numeric expression” box > From the key pad click “+” > Select “k3” and push it into the “Numeric expression” box > From the key pad click “+” > Select “k4” and push it into the “Numeric expression” box > OK (Fig 6.15)

You will achieve the same results.

Figure 6.15



Target Variable: t_know

Numeric Expression: k1 + k2 + k3 + k4

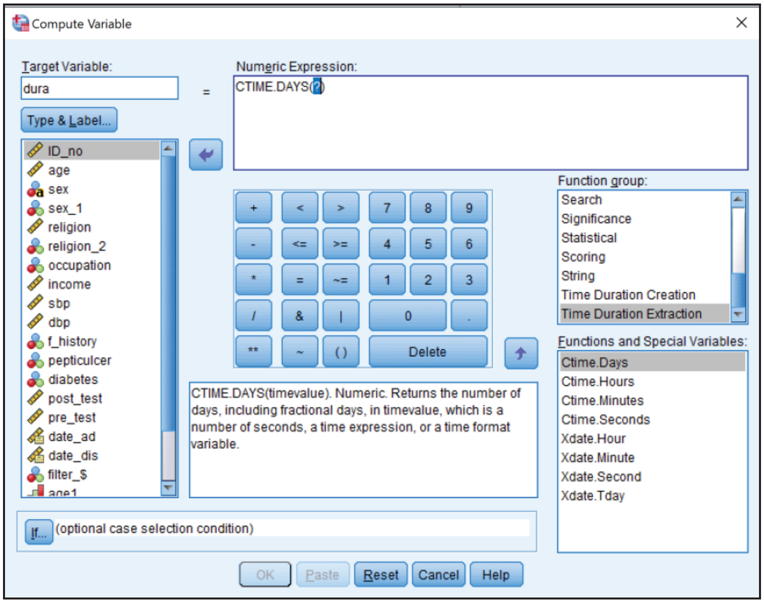
6.7 Calculation of duration

SPSS can extract time duration from dates. For instance, you have the data on date of admission (variable name is date_ad) and date of discharge (date_dis) of patients admitted to a hospital (use the data file **Data_3.sav**). Now, you want to calculate the duration of hospital stay (date of discharge minus date of admission). SPSS can calculate this for you. Use the following commands.

Transform > Compute Variable > Type “dura” under “Target Variable” > Click on “Time Duration Extraction” in the “Function Group” box > From “Functions and special variables” box select “Ctime.Days” > Click on the up arrow (at the left side of the box). You will see CTIME.DAYS(?) appears in the “Numeric Expression” box > Select “date_dis” from “Type and Label” box > Push it into the “Numeric Expression” box > Click on – (minus sign from the pad) > Select “date_ad” from “Type and Label” box and push it into the “Numeric Expression” box > OK (Fig 6.16)

You will notice that SPSS has generated a new variable “dura” (the last variable) that contains the duration of hospital stay of each subject in the dataset. You can check the data by making a frequency distribution table of the variable “dura”.

Figure 6.16



Target Variable: dura

Numeric Expression: CTIME.DAYS(?)

CTIME.DAYS(timevalue). Numeric. Returns the number of days, including fractional days, in timevalue, which is a number of seconds, a time expression, or a time format variable.

6.8 Selecting a sub-group for analysis

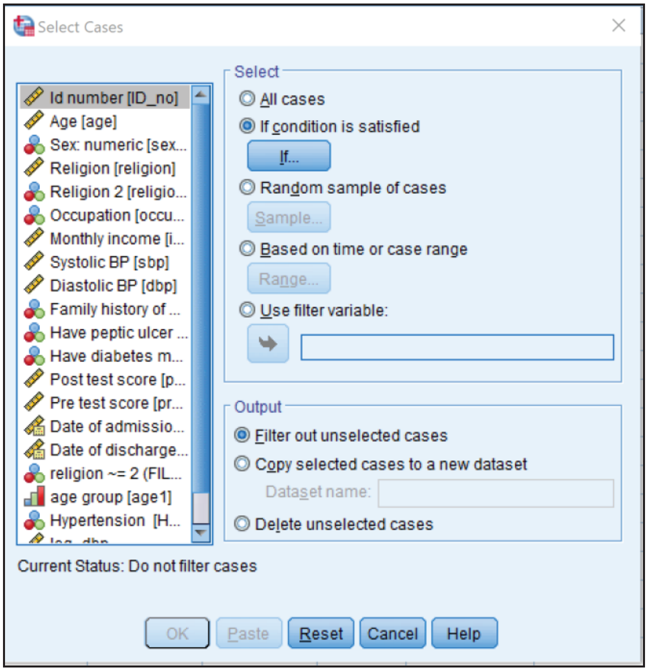
You can select a specific sub-group for the analysis of your data. Suppose you want to analyze your data only for those who have diabetes mellitus. In the dataset, the variable “diabetes” is coded as “1= yes (have diabetes)” and “2= no (do not have diabetes)”. To select the group who have diabetes (i.e., diabetes=1), use the following commands.

Data > Select “Select cases” > Select “If condition is satisfied” > Click on “If” > Select the variable “diabetes” and push it into the empty box > Click “=” and then “1” from the number pad > Continue > OK (Figs 6.17 and 6.18)

This will exclude the subjects who do not have diabetes from the subsequent analysis. The analysis will be only for those who have diabetes. If you generate a frequency distribution table for “sex”, you will see that n= 45 (Table 6.6). Now, to get back to all the subjects for analysis (i.e., to deselect the subgroup), use the commands:

Data > Select cases > Select "All cases" > OK (Fig 6.17)

Figure 6.17



Select Cases

Select

○ All cases

● If condition is satisfied

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

Output

● Filter out unselected cases

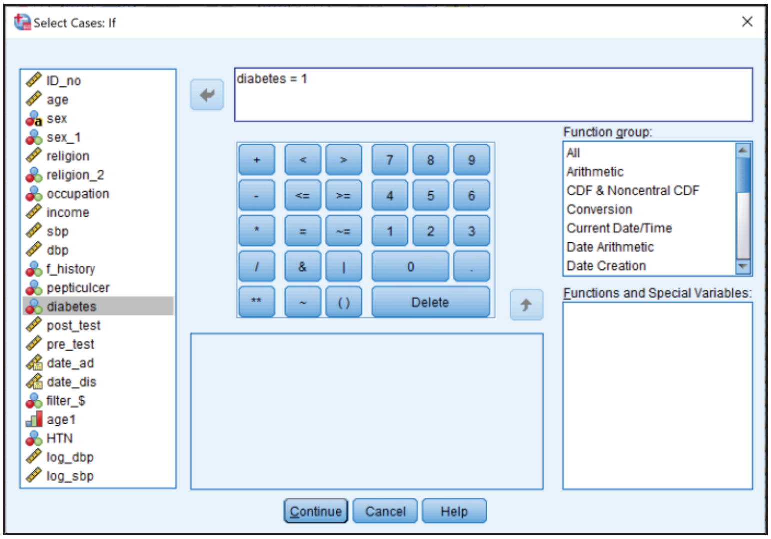
○ Copy selected cases to a new dataset

Dataset name:

○ Delete unselected cases

Current Status: Do not filter cases

Figure 6.18



Select Cases: If

```
diabetes = 1
```

Table 6.6 Distribution of sex among diabetic patients

Sex: string					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	20	44.4	44.4	44.4
	male	25	55.6	55.6	100.0
	Total	45	100.0	100.0	

6.9 Split file

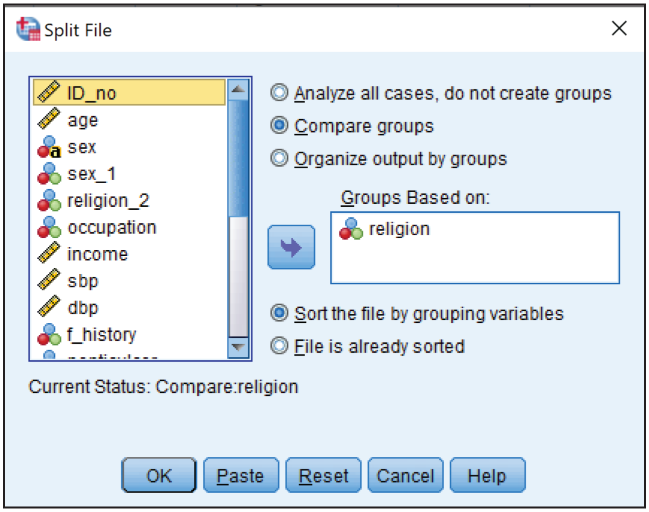
This function will split the file by the categories of the variable(s) selected. If this option is selected, SPSS will process data and produce the outputs disaggregated by the category of the variable selected. If you want to get the outputs disaggregated by religion, use the following commands.

Data > Split File > Select "Compare groups" (you can also select the option "Organize output by groups" instead) > Select "religion" and push it into "Groups based on" box > OK (Fig 6.19)

Now, make a frequency distribution of the variable sex. You will see that the distribution of sex is provided by religion (Table 6.7). To revert the function, just use the following commands.

Data > Split File > Select "Analyze all cases, do not create groups" > OK (Fig 6.19)

Figure 6.19



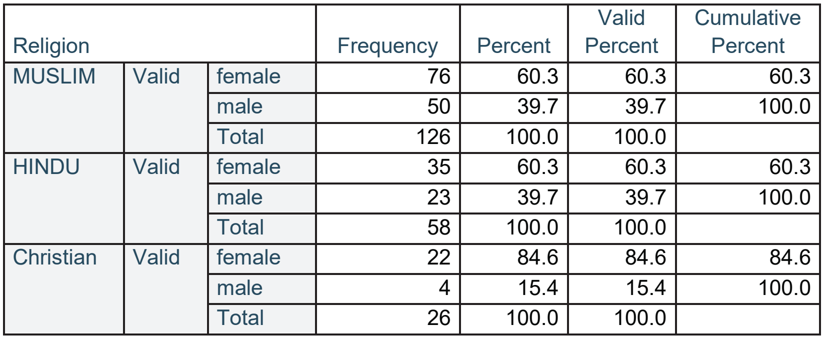
○ Analyze all cases, do not create groups
● Compare groups
○ Organize output by groups

Groups Based on: religion

● Sort the file by grouping variables
○ File is already sorted

Current Status: Compare:religion

Table 6.7 Distribution of sex by religion



Religion			Frequency	Percent	Valid Percent	Cumulative Percent
MUSLIM	Valid	female	76	60.3	60.3	60.3
		male	50	39.7	39.7	100.0
		Total	126	100.0	100.0	
HINDU	Valid	female	35	60.3	60.3	60.3
		male	23	39.7	39.7	100.0
		Total	58	100.0	100.0	
Christian	Valid	female	22	84.6	84.6	84.6
		male	4	15.4	15.4	100.0
		Total	26	100.0	100.0	

6.10 Creating variable sets

Typically, we have many variables in our dataset. But we may require only a few variables for a specific analysis (e.g., logistic regression, multiple linear regression). During analysis, we need to find them out from the list of all variables in the dataset. For easy access to the variables needed for an analysis, we can create sets with the variables of interest.

Suppose you want to do a logistic regression analysis with the variables - age, sex, religion, family history of diabetes (variable name is f_history), and diabetes. To make a set with these variables, you need to select a name for the set. Let's consider the name of the set as "LR_1". Now, to make the set (LR_1) with these variables, use the following commands.

Utilities > Define variable sets (Fig 6.20) > Write "LR_1" in the box "Set name" > Select the variables "age, sex, religion, f_history and diabetes" and push them into the "Variables in set" box > Click "Add set" > Close (Fig 6.21)

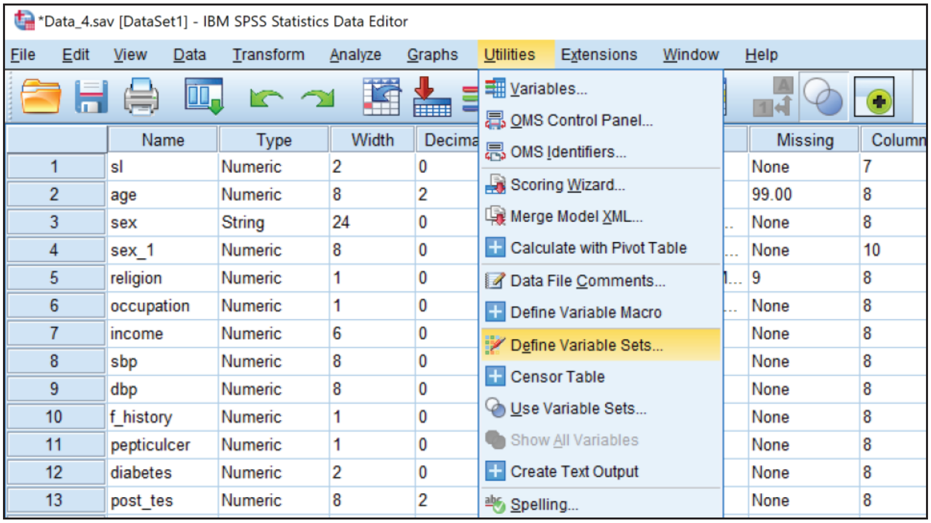
This will generate a set "LR_1" with the variables as mentioned above. Note that you cannot see anything in the data view or variable view. To see and activate the set, use the following commands.

Utilities > Use variable sets > From the template (Fig 6.22) deselect "ALLVARIABLES" and "NEWVARIABLES" > Select "LR_1" > OK

You will notice that the variables included in the set "LR_1" are now visible in the data editor. You can do any analysis with these variables. To go back to the dataset with all the variables, use the following commands.

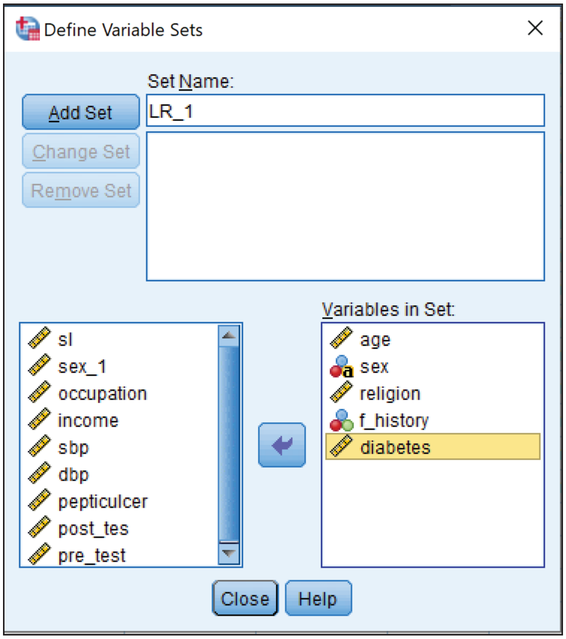
Utilities > Click on "Show all variables"

Figure 6.20



	Name	Type	Width	Decima	Missing	Column
1	sl	Numeric	2	0	None	7
2	age	Numeric	8	2	99.00	8
3	sex	String	24	0	None	8
4	sex_1	Numeric	8	0	None	10
5	religion	Numeric	1	0	9	8
6	occupation	Numeric	1	0	None	8
7	income	Numeric	6	0	None	8
8	sbp	Numeric	8	0	None	8
9	dbp	Numeric	8	0	None	8
10	f_history	Numeric	1	0	None	8
11	pepticulcer	Numeric	1	0	None	8
12	diabetes	Numeric	2	0	None	8
13	post_tes	Numeric	8	2	None	8

Figure 6.21



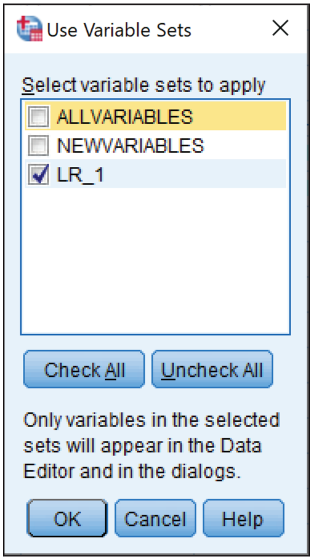
Set Name: LR_1

sl
sex_1
occupation
income
sbp
dbp
pepticulcer
post_tes
pre_test

Variables in Set:

age
sex
religion
f_history
diabetes

Figure 6.22



Select variable sets to apply

☐ ALLVARIABLES
☐ NEWVARIABLES
☑ LR_1

Only variables in the selected sets will appear in the Data Editor and in the dialogs.

7

Testing of Hypothesis

The current and following chapters provide basic information on how to select statistical tests for testing hypotheses, perform the statistical tests by SPSS and interpret the results of common problems related to health and social sciences research. Before we proceed, let us discuss a little bit about the hypothesis.

A hypothesis is a statement about one or more populations. The hypothesis is usually concerned about the parameter of the population about which the statement is made. There are two types of statistical hypothesis, Null (H_0) and Alternative (H_A) hypothesis. The null hypothesis is the hypothesis of equality or no difference. The null hypothesis always says that the two or more quantities (parameters) are equal. *Note that, we always test the null hypothesis, not the alternative hypothesis*. Using a statistical test, we either reject or do not reject the null hypothesis. If we can reject the null hypothesis, then only we can accept the alternative hypothesis. It is, therefore, necessary to have a very clear understanding of the null hypothesis.

Suppose we are interested in determining the association between coffee drinking and stomach cancer. In this situation, the null hypothesis is "there is no association between coffee drinking and stomach cancer (or, coffee drinking and stomach cancer are independent)", while the alternative hypothesis is "there is an association between coffee drinking and stomach cancer (or, coffee drinking and stomach cancer are not independent) ". If we can reject the null hypothesis by a statistical test (i.e., if the test is significant; p-value <0.05), then only we can conclude that there is an association between coffee drinking and stomach cancer.

Various statistical tests are available to test a hypothesis. Selecting an appropriate statistical test is the key to effective data analysis. The statistical test used to test a hypothesis depends on the study design, data type, distribution of data, and objective of the study. It is, therefore, important to understand the nature of the variable (categorical or quantitative), measurement type (nominal, ordinal, interval or ratio scale), as well as the study design. Following table (Table 7) provides basic guidelines about the selection of statistical tests depending on the type of data and situation.

Table 7. Selecting statistical test for hypothesis testing

7.1 Association between quantitative and categorical or quantitative variables

	Situation for hypothesis testing	Data normally distributed	Data non-normal
1	**Comparison with a single population mean (with a fixed value)** Example: You have taken a random sample from a population of diabetic patients to assess the mean age. Now, you want to test the hypothesis of whether the mean age of diabetic patients in the population is 55 years or not.	1-sample t-test	Sign test/ Wilcoxon Signed Rank test
2	**Comparison of means of two related samples** Example: You want to test the hypothesis whether the drug "Inderal" is effective in reducing blood pressure (BP) or not. To test the hypothesis, you have selected a group of subjects and measured their BP before administration of the drug (measurements before treatment; or pre-test). Then you have given the drug "Inderal" to all the subjects and measured their BP after one hour (measurements after treatment; or post-test). Now you want to compare if the mean BP before (pre-test) and after (post-test) administration of the drug is same or not.	Paired t-test	Sign test/ Wilcoxon Signed Rank test
3	**Comparison between two independent sample means [association between a quantitative and a categorical/qualitative variable with *2 levels*]** Example: You have taken a random sample of students from a university. Now, you want to test the hypothesis if the mean systolic blood	Independent samples t-test	Mann-Whitney U test (also called Wilcoxon Rank Sum test)

	Situation for hypothesis testing	Data normally distributed	Data non-normal
	pressure of male and female students is same or not.		
4	**Comparison of more than two independent sample means [association between a quantitative and a categorical variable with *more than 2 levels*]** Example: You have taken a random sample from a population. You want to test the hypothesis if the mean income of different religious groups (e.g., Muslim, Hindu and Christian) is same or not. Another example, you have three drugs, A, B and C. You want to investigate whether all these three drugs are equally effective in reducing the blood pressure or not.	One way ANOVA	Kruskal Wallis test
5	**Association between two quantitative variables** Example: You want to test the hypothesis if there is a correlation between systolic BP and age.	Pearson's correlation	Spearman's correlation (Also valid for ordinal qualitative data)
6	**Association between a quantitative and an ordinal variable** Example: You want to test the hypothesis if there is a correlation between systolic BP and severity of anaemia (no, mild, moderate, severe).	Spearman's correlation, if the ordinal variable has 5 or more levels. Otherwise, use Kendall's Tau-B statistics	
7	**Association between two ordinal variables** Example: You want to test the hypothesis if there is a correlation between severity of pain and stage of cancer.	Spearman's correlation, if both the ordinal variables have 5 or more levels. Otherwise, use Kendall's Tau-B statistics	

7.2 Association between two categorical variables

	Situation for hypothesis testing	Test statistics
1	**Association between two categorical variables (independent samples)** Example: You have taken a random sample from a population and want to test the hypothesis if there is an association between sex and asthma. Another example, you want to assess the association between smoking and stomach cancer.	Chi-square test/ Fisher's Exact test
2	**Association between two categorical variables of related samples, such as data of a matched case-control study design** Example: You want to test the hypothesis if there is an association between diabetes mellitus and heart disease, when the data is matched for smoking or other variables (a matched case-control study design).	McNemar test

7.3 Multivariable analysis

	Type of outcome/dependent variable	Type of multivariable analysis
1	Outcome variable (also called dependent variable) is in interval or ratio scale – e.g., blood pressure, birth weight, blood sugar.	Multiple linear regression; Analysis of variance (ANOVA)
2	Dependent variable is a dichotomous categorical variable (i.e., a nominal categorical variable with two levels) – e.g., disease (present or absent); ANC (taken or not taken); treatment outcome (cured or not cured).	Multiple logistic regression
3	Dependent variable is a nominal categorical variable with more than two levels – e.g., treatment seeking behaviour (e.g., did not receive treatment, received homeopathic treatment, received allopathic treatment); cause of death (cancer, heart disease, pneumonia).	Multi-nominal logistic regression

	Type of outcome/dependent variable	**Type of multivariable analysis**
4	Dependent variable is an ordinal categorical variable – e.g., severity of anaemia (no anaemia, mild to moderate anaemia, severe anaemia); stage of cancer (stage1, stage 2, stage 3); severity of pain (mild, moderate, severe).	Proportional odds regression (Ordinal regression)
5	Dependent variable is time-to-outcome (e.g., time -to-death, time-to-recurrence, time-to-cure).	Proportional hazards analysis (Cox regression)
6	Dependent variable is counts – e.g., number of post-operative infections; number of patients admitted with heart disease in a hospital, number of road traffic accidents treated in the emergency department.	Poisson regression
7	Incidence rates – incidence rate of tuberculosis; incidence rate of pneumonia; incidence rate of car accidents.	Poisson regression

7.4 Agreement analysis

	Situation for hypothesis testing	**Test statistics**
1	**Agreement between two quantitative variables** Example: You want to test the hypothesis if two methods of blood sugar measurements agree with each other.	Bland Altman test/plots
2	**Agreement between two categorical variables** Example: You want to test the hypothesis if diagnosis of cataract agrees between two physicians .	Kappa estimates

8

Student's t-test for Hypothesis Testing

Student's t-test is commonly known as t-test. It is a commonly used parametric statistical method to test a hypothesis. There are several types of t-tests used in different situations (Table 7 of chapter 7), such as a) One-sample t-test; b) Independent samples t-test; and c) Paired t-test. In this chapter, we shall discuss all these t-tests and interpretation of the results. Use the data file <**Data_3.sav**> for practice.

8.1 One-sample t-test

One-sample t-test is done to compare the mean of a variable with a hypothetical value. For example, we have collected data on diastolic BP (variable name: dbp) of students of the State University of Bangladesh taking a random sample. We are interested to know if the mean diastolic BP of the students is 80 mmHg or not.

Hypothesis

Null hypothesis (H_0): The mean diastolic BP of students is equal to 80 mmHg in the population (study population is the students of State University of Bangladesh).

Alternative hypothesis (H_A): The mean diastolic BP of students is different from (not equal to) 80 mmHg in the population.

Assumptions

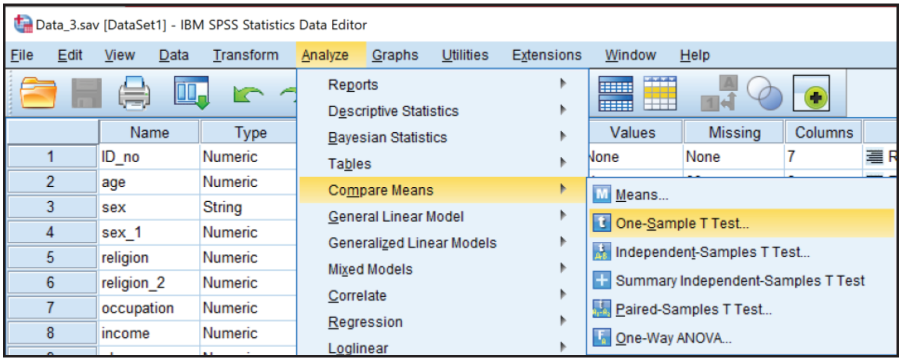
1. The distribution of diastolic BP in the population is normal;
2. The sample is a random sample from the population.

The first job, before hypothesis testing, is to check whether the distribution of diastolic BP is normal in the population (assumption 1). To do this, check the histogram and/or Q-Q plot of diastolic BP and do the formal statistical test of normality (K-S test or Shapi-

-ro Wilk test) as discussed in chapter 5. If the assumption is met (diastolic BP is at least approximately normal), do the 1-sample t-test, otherwise, we have to use the non-parametric test (discussed in chapter 18). Assume that the diastolic BP is normally distributed in the population. Use the following commands to do the 1-sample t-test.

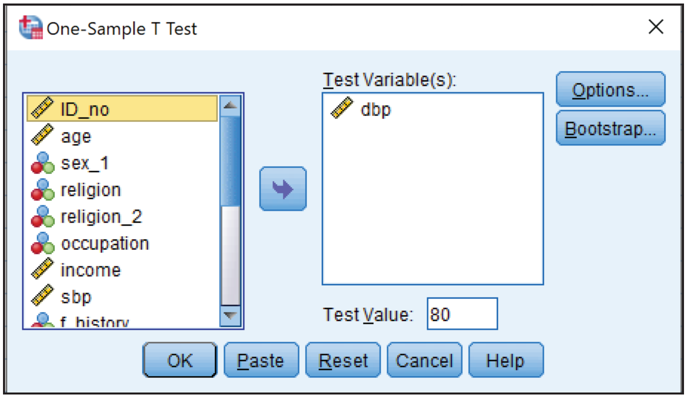
> Analyze > Compare means > One-sample t-test > Select the variable "dbp" and push it into the "Test variable(s)" box > Click in the "Test value" box and write "80" > OK (Fig 8.1 and 8.2)

Figure 8.1



	Name	Type
1	ID_no	Numeric
2	age	Numeric
3	sex	String
4	sex_1	Numeric
5	religion	Numeric
6	religion_2	Numeric
7	occupation	Numeric
8	income	Numeric

Figure 8.2



Test Variable(s): dbp

Test Value: 80

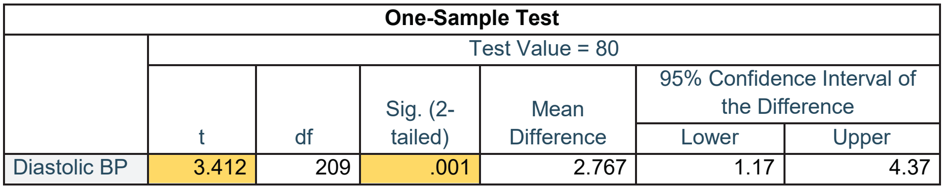
8.1.1 Outputs

The SPSS will provide the following tables (Tables 8.1 and 8.2).

Table 8.1 Descriptive statistics of diastolic BP

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
Diastolic BP	210	82.77	11.749	.811

Table 8.2 One-sample t-test results



One-Sample Test

	Test Value = 80					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Diastolic BP	3.412	209	.001	2.767	1.17	4.37

8.1.2 Interpretation

In this example, we have tested the *null hypothesis* "the mean diastolic BP of the students is equal to 80 mmHg in the population". Data shows that the mean diastolic BP of the sample of the students is 82.77 mmHg with an SD of 11.7 mmHg (Table 8.1). One-sample t-test results are shown in Table 8.2. The calculated value of "t" is 3.41 and the p-value (Sig. 2-tailed) is 0.001. Since the p-value is <0.05, we can reject the null hypothesis at 95% confidence level. This means that the mean diastolic BP of the students (in the population) from where the sample is drawn is different from 80 mmHg (p<0.001). The SPSS has also provided the difference between the observed value (82.77) and hypothetical value (80.0) as the mean difference (which is 2.76) and its 95% confidence interval (1.17 – 4.37) (Table 8.2).

8.2 Independent samples t-test

The independent samples t-test involves one categorical variable with two levels (categories) and one quantitative variable. This test is done to compare the means of two categories of the categorical variable.

For example, we are interested to investigate if the mean age of diabetic and non-diabetic patients, in the population, is same or not. Here, the test variable (dependent variable) is age (quantitative) and the categorical variable is diabetes, which has two levels/categories (have diabetes and do not have diabetes). Before doing the test, we need to check the assumption 1 [i.e., age is normally distributed at each level of the independent variable (diabetes)].

Hypothesis

H_0: The mean age of the diabetic and non-diabetic patients is same in the population.

H_A: The mean age of the diabetic and non-diabetic patients is different (not same) in the population.

Assumptions

1. The dependent variable (age) is normally distributed at each level of the independent (diabetes) variable;
2. The variances of the dependent variable (age) at each level of the independent variable (diabetes) are same/equal;
3. Subjects represent random samples from the populations.

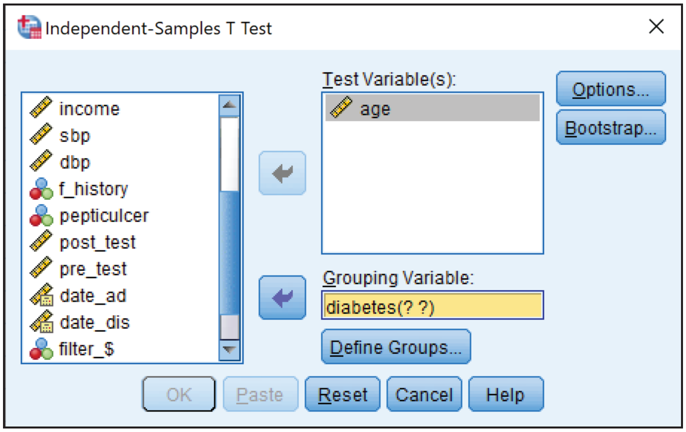
8.2.1 Commands

Use the following commands to do the independent samples t-test. Before doing the test, we need to remember/check (from codebook or variable view) the category code numbers of diabetes. In our example, we have used code "1" for defining "have diabetes" and "2" for "do not have diabetes".

> Analyze > Compare means > Independent samples t-test (Fig 8.1) > Select "age" and push it into the "test variable(s)" box and select "diabetes" for "grouping variable" box (Fig 8.3) > Click on "define groups" > Type 1 in "Group 1" box and type 2 in "Group 2" box (Fig 8.4) > Continue > OK

Note: You will need to use exactly the same code numbers as it is in the dataset for the grouping variable. Otherwise, SPSS cannot analyse the data.

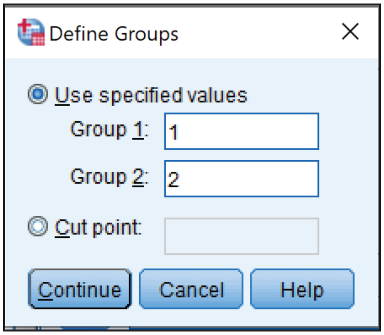
Figure 8.3



Test Variable(s): age

Grouping Variable: diabetes(? ?)

Figure 8.4



Define Groups

● Use specified values

Group 1: 1

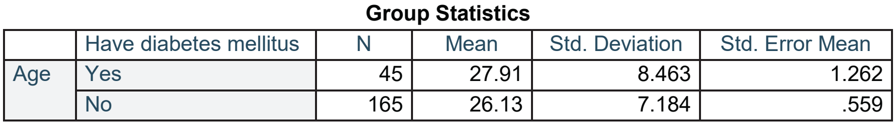
Group 2: 2

○ Cut point:

8.2.2 Outputs

The SPSS will produce the outputs as shown in Tables 8.3 and 8.4.

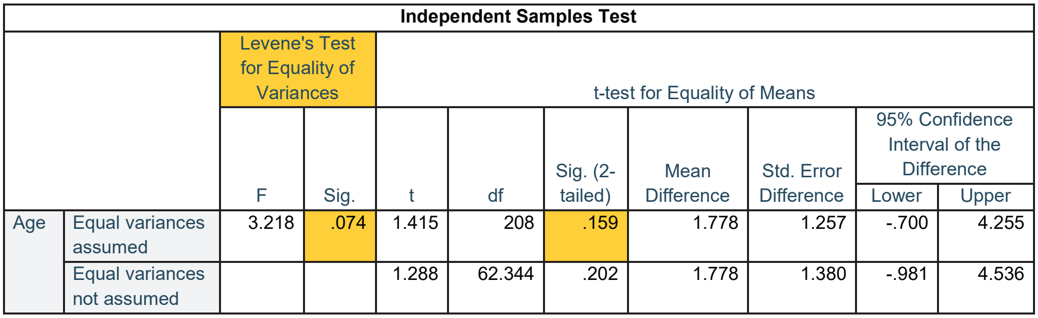
Table 8.3 Descriptive statistics of age by grouping variable (having diabetes)



Group Statistics

	Have diabetes mellitus	N	Mean	Std. Deviation	Std. Error Mean
Age	Yes	45	27.91	8.463	1.262
	No	165	26.13	7.184	.559

Table 8.4 Independent samples t-test results



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age	Equal variances assumed	3.218	.074	1.415	208	.159	1.778	1.257	-.700	4.255
	Equal variances not assumed			1.288	62.344	.202	1.778	1.380	-.981	4.536

8.2.3 Interpretation

Table 8.3 shows the descriptive measures of age by the grouping variable (diabetes). We can see that there are 45 persons with diabetes and 165 persons without diabetes. The mean age of the diabetic persons is 27.9 (SD 8.46) years and that of the non-diabetic persons is 26.1 (SD 7.18) years.

Table 8.4 shows the independent samples t-test results. The first portion of the table indicates the *Levene's test* results. This test is done to understand if the variances of age in the two categories of diabetes are homogeneous (equal) or not (assumption 2). Look at the p-value (Sig.) of the Levene's test, which is 0.074. Since the p-value is >0.05, it indicates that the variances of age of the diabetic and non-diabetic persons are homogeneous (assumption 2 is fulfilled).

Now, look at the other portion of the table, the *t-test for equality of means*. Here, we must decide which p-value we shall consider. If the Levene's test p-value is >0.05, take the t-test results at the upper row, i.e., t-test for "Equal variances assumed". If the Levene's test p-value is ≤0.05, take the t-test results at the lower row, i.e., t-test for "Equal variances not assumed".

In this example, as the Levene's test p-value is >0.05, we shall consider the t-test results of "Equal variances assumed", i.e., the upper row. Table 8.4 shows that the calculated t-value is 1.415, and the p-value (2-tailed) is 0.159 (which is >0.05) with 208 degrees of freedom. We cannot, therefore, reject the null hypothesis. This means that the mean age of diabetic and non-diabetic persons in the population from where samples are drawn is not different (p=0.159).

8.3 Paired t-test

The paired t-test is done to compare the difference between two means of *related* samples. Related samples indicate measurements taken from the same subjects in two or more different times/situations. For example, you have organized a training for 32 staff of your organization. To evaluate the effectiveness of the training, you have taken a pre-test before the training to assess the current knowledge of the participants. At the end of the training, you have again taken an examination (post-test). Now you want to compare if the training has increased their knowledge or not. Another example is "To determine the effectiveness of a drug (e.g., Inderal) in reducing the systolic blood pressure (BP), you have selected a random sample from a population. You have measured the systolic BP of all the individuals before giving the drug (pre-test or baseline). You have again measured their systolic BP one-hour after giving the drug (post-test or endline)". Paired t-test is the appropriate test to compare the means in both the situations.

Hypothesis

H_0: There is no difference in the mean scores before and after the training (for example 1).

H_A: The mean scores are different before and after the training.

Assumptions

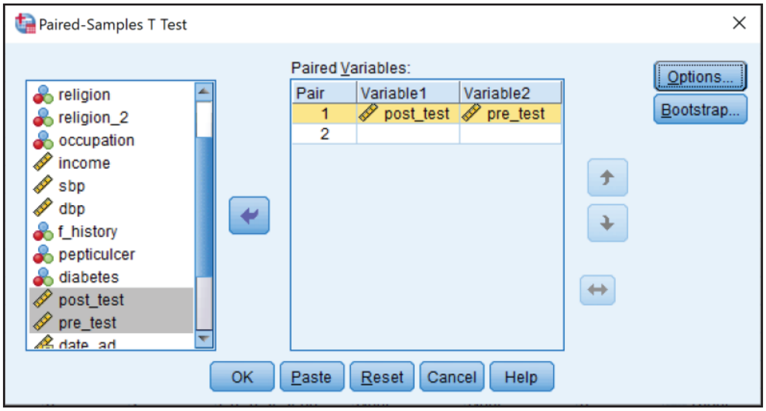
1. The difference between two measurements (pre- and post-test) of the dependent variable (examination scores) is normally distributed;

2. Subjects represent a random sample from the population.

8.3.1 Commands

Analyze > Compare means > Paired-samples t-test > Select the variables "post-test" and "pre-test" and push them into the "Paired variables" box > OK (Fig 8.5)

Figure 8.5



Paired-Samples T Test

religion
religion_2
occupation
income
sbp
dbp
f_history
pepticulcer
diabetes
post_test
pre_test
date_ad

Paired Variables:

Pair	Variable1	Variable2
1	post_test	pre_test
2		

8.3.2 Outputs

The SPSS will produce the following outputs (Tables 8.5 to 8.7).

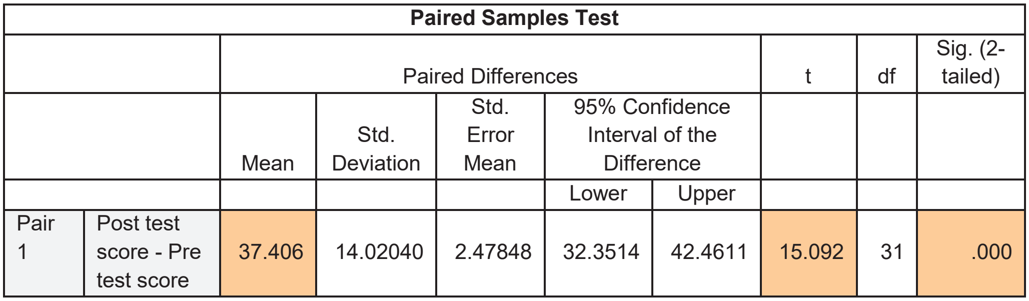
Table 8.5 Descriptive statistics of pre- and post-test scores

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Post test score	90.9844	32	8.44096	1.49216
	Pre test score	53.5781	32	15.42835	2.72737

Table 8.6 Correlation between pre- and post-test scores

Paired Samples Correlations				
		N	Correlation	Sig.
Pair 1	Post test score & Pre test score	32	.433	.013

Table 8.7 Paired samples t-test results



Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Post test score - Pre test score	37.406	14.02040	2.47848	32.3514	42.4611	15.092	31	.000

8.3.3 Interpretation

Table 8.5 shows the means of both the pre- (53.5) and post-test (90.9) scores along with the standard deviations and standard errors. Looking at the mean scores, we can form an impression on whether the training has increased the mean score or not. We can see that the post-test mean is 90.9, while the pre-test mean is 53.5. To understand, if the difference between post-test mean and pre-test mean is significant or not, we have to check the paired samples t-test results (Table 8.7). Table 8.7 shows that the mean difference between the post- and pre-test scores is 37.4. The calculated t-value is 15.09 and the p-value (sig.) is 0.000. As the p-value is <0.05, reject the null hypothesis. This indicates that the mean knowledge score has been increased significantly after the training ($p<0.001$). Note that for conclusion, we do not need Table 8.6.

9

Analysis of Variance (ANOVA)

Analysis of variance or ANOVA is a commonly used statistical method for testing a hypothesis. ANOVA is used to compare the means when the categorical independent variable has more than 2 levels. There are several types of ANOVA tests, such as one-way ANOVA, two-way ANOVA, repeated-measures ANOVA and others. In this chapter, one-way and two-way ANOVA are discussed. The repeated measures ANOVA is discussed in chapter 10. Use the data file <**Data_3.sav**> for practice.

9.1 One-way ANOVA

The one way-ANOVA test is used to compare the means of more than two groups, while the t-test compares the means of two groups. The one-way ANOVA test involves two variables, one categorical variable with more than two levels/categories (for example, in our data the variable "religion" [variable name "religion_2"] has 4 categories – Muslim, Hindu, Christian and Buddhism) and a quantitative variable (e.g., income, age, blood pressure). Suppose you want to assess if the mean income (SPSS variable name is "income") of all the religious groups is same or not in the population. The one-way ANOVA is the appropriate test for this, if the assumptions are met.

Hypothesis

(H_0): The mean income of all the religious groups is same/equal.

(H_A): Not all the means (of income) of religious groups are same.

Assumptions

1. The dependent variable (income) is normally distributed at each level of the independent variable (religion);
2. The variances of the dependent variable (income) for each level of the independ-

ent variable (religion) are same (homogeneity of variances); and

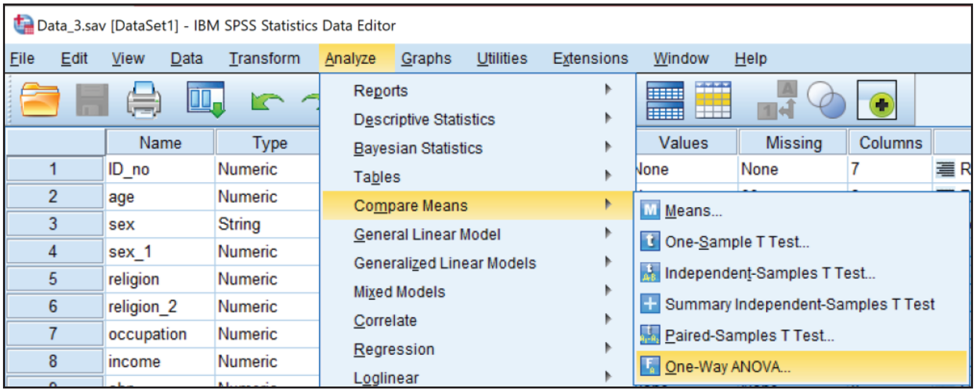
3. Subjects represent random samples from the populations.

If the variances of the dependent variable in all the categories are not equal (violation of assumption 2), but sample size in all the groups is large and similar, ANOVA can be used.

9.1.1 Commands

Analyze > Compare means > One-way ANOVA (Fig 9.1) > Select "income" and push it into the "Dependent list" box > Select "religion_2" for the "Factor" box (Fig 9.2) > Options > Select "Descriptive" and "Homogeneity of variance test" > Continue > OK

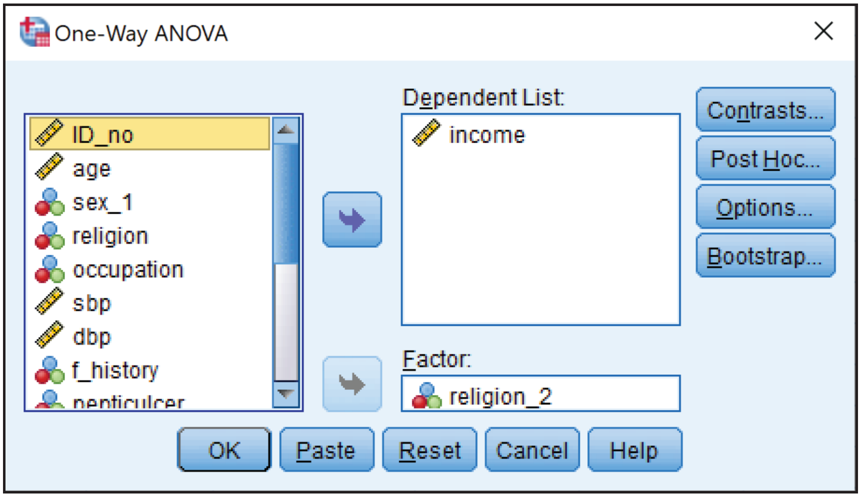
Figure 9.1



Data_3.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Means...
One-Sample T Test...
Independent-Samples T Test...
Summary Independent-Samples T Test
Paired-Samples T Test...
One-Way ANOVA...
Values
Missing
Columns

	Name	Type
1	ID_no	Numeric
2	age	Numeric
3	sex	String
4	sex_1	Numeric
5	religion	Numeric
6	religion_2	Numeric
7	occupation	Numeric
8	income	Numeric

Figure 9.2



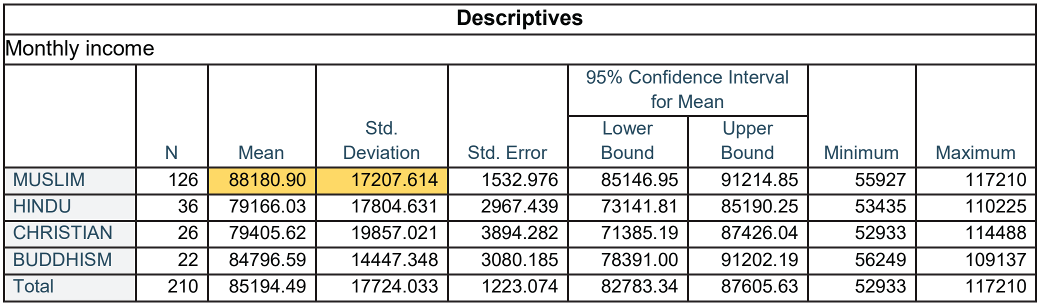
Dependent List: income

Factor: religion_2

9.1.2 Outputs

The SPSS will generate the following outputs (Table 9.1-9.3).

Table 9.1 Descriptive statistics of income by religious groups

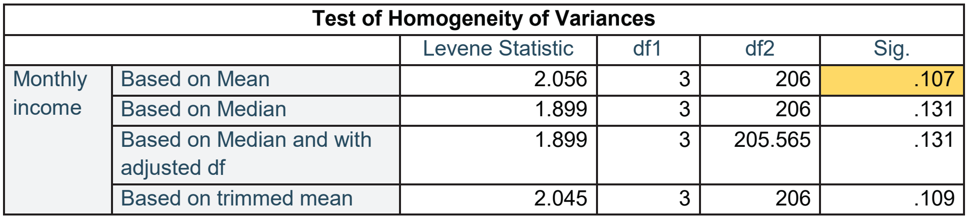


Descriptives

Monthly income

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
MUSLIM	126	88180.90	17207.614	1532.976	85146.95	91214.85	55927	117210
HINDU	36	79166.03	17804.631	2967.439	73141.81	85190.25	53435	110225
CHRISTIAN	26	79405.62	19857.021	3894.282	71385.19	87426.04	52933	114488
BUDDHISM	22	84796.59	14447.348	3080.185	78391.00	91202.19	56249	109137
Total	210	85194.49	17724.033	1223.074	82783.34	87605.63	52933	117210

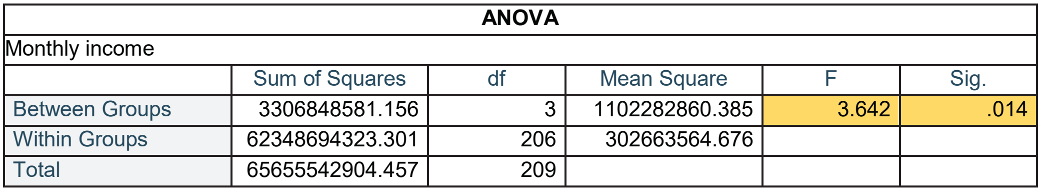
Table 9.2 Levene's test for homogeneity of variances of income in different religious groups



Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Monthly income	Based on Mean	2.056	3	206	.107
	Based on Median	1.899	3	206	.131
	Based on Median and with adjusted df	1.899	3	205.565	.131
	Based on trimmed mean	2.045	3	206	.109

Table 9.3 ANOVA test results



ANOVA

Monthly income

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3306848581.156	3	1102282860.385	3.642	.014
Within Groups	62348694323.301	206	302663564.676		
Total	65655542904.457	209			

9.1.3 Interpretation

In this example, we have used "income" as the dependent variable and "religion" as the independent (or factor) variable. The independent variable (religion) has 4 categories (levels) – Muslim, Hindu, Christian and Buddhism.

Table 9.1 provides all the descriptive measures (mean, SD, SE, 95% CI) of income by religion. For example, the mean income of Muslims is 88,180.9 with an SD of 17,207.6.

The second table (Table 9.2) shows the test results of homogeneity of variances (Levene's test). This test was undertaken to understand if all the group variances of income are equal (assumption 2). Table 9.2 provides the Levene's test results for both the mean and median. Consider the test results for the mean. The p-value (Sig.) of the Levene's test is

0.107. Since the p-value is >0.05, the variances of income in all the religious groups are equal (i.e., assumption 2 is not violated).

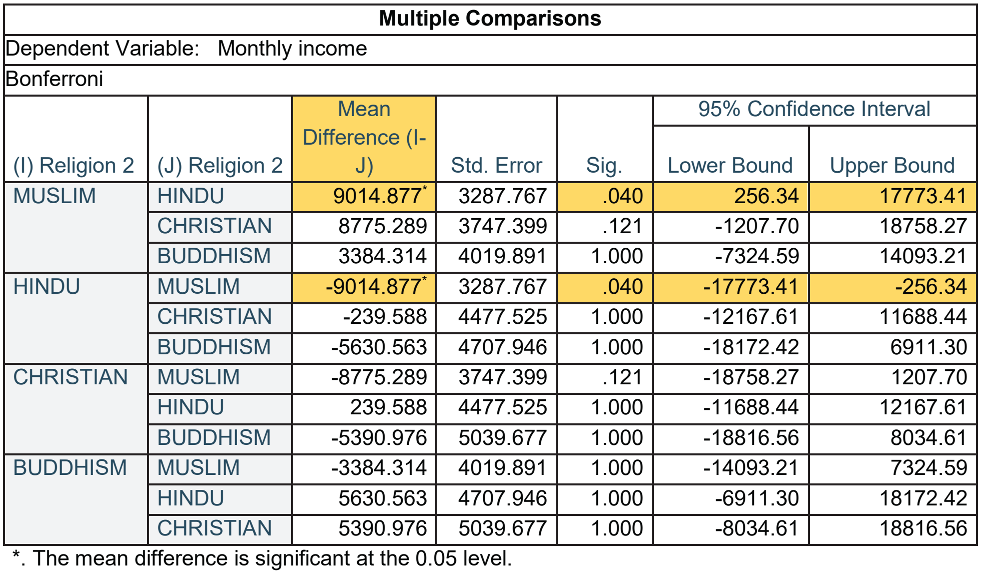
Now, look at the ANOVA table (Table 9.3). The value of the F-statistic is 3.642 and the p-value is 0.014. Since the p-value is <0.05, reject the null hypothesis. *This means that, not all group means (of income) are same.*

However, the ANOVA test does not provide information about which group means are different. To understand which group means are different, we need to use the *post hoc multiple comparison test,* such as *Tukey's test or Bonferroni test*. Use the following commands to get the post hoc test results. *Note that if the ANOVA test (F-test) is not significant (i.e., p-value is >0.05), we do not need the post-hoc test.*

> Analyze > Compare means > One-way ANOVA > Select "income" and push it into the "Dependent list" box > Select "religion_2" for the "Factor" box > Options > Select "Descriptive", and "Homogeneity of variance test" > Continue > Post Hoc > Select "Bonferroni" (or Tukey) > Continue > OK

The SPSS will produce the following table (Table 9.4) in addition to others.

Table 9.4 Comparisons of mean income between the religious groups (Bonferroni's test results)



Multiple Comparisons

Dependent Variable: Monthly income

Bonferroni

(I) Religion 2	(J) Religion 2	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
MUSLIM	HINDU	9014.877*	3287.767	.040	256.34	17773.41
	CHRISTIAN	8775.289	3747.399	.121	-1207.70	18758.27
	BUDDHISM	3384.314	4019.891	1.000	-7324.59	14093.21
HINDU	MUSLIM	-9014.877*	3287.767	.040	-17773.41	-256.34
	CHRISTIAN	-239.588	4477.525	1.000	-12167.61	11688.44
	BUDDHISM	-5630.563	4707.946	1.000	-18172.42	6911.30
CHRISTIAN	MUSLIM	-8775.289	3747.399	.121	-18758.27	1207.70
	HINDU	239.588	4477.525	1.000	-11688.44	12167.61
	BUDDHISM	-5390.976	5039.677	1.000	-18816.56	8034.61
BUDDHISM	MUSLIM	-3384.314	4019.891	1.000	-14093.21	7324.59
	HINDU	5630.563	4707.946	1.000	-6911.30	18172.42
	CHRISTIAN	5390.976	5039.677	1.000	-8034.61	18816.56

*. The mean difference is significant at the 0.05 level.

9.1.3.1 Interpretation of multiple comparisons table

Table 9.4 shows the mean difference in income of different religious groups. We can see that the difference in mean income of Muslims and Hindus is 9,014.87 (the minus sign in row 4 indicates that Muslims have a greater income than Hindus). The p-value (Sig.) of this difference is 0.040, which is <0.05. This indicates that the mean income of Muslims and Hindus may be different in the population (Muslims have a higher mean income than the Hindus). The difference of means of other religious groups is not significant, as the p-values are >0.05. The table has also provided the 95% CI of the mean differences.

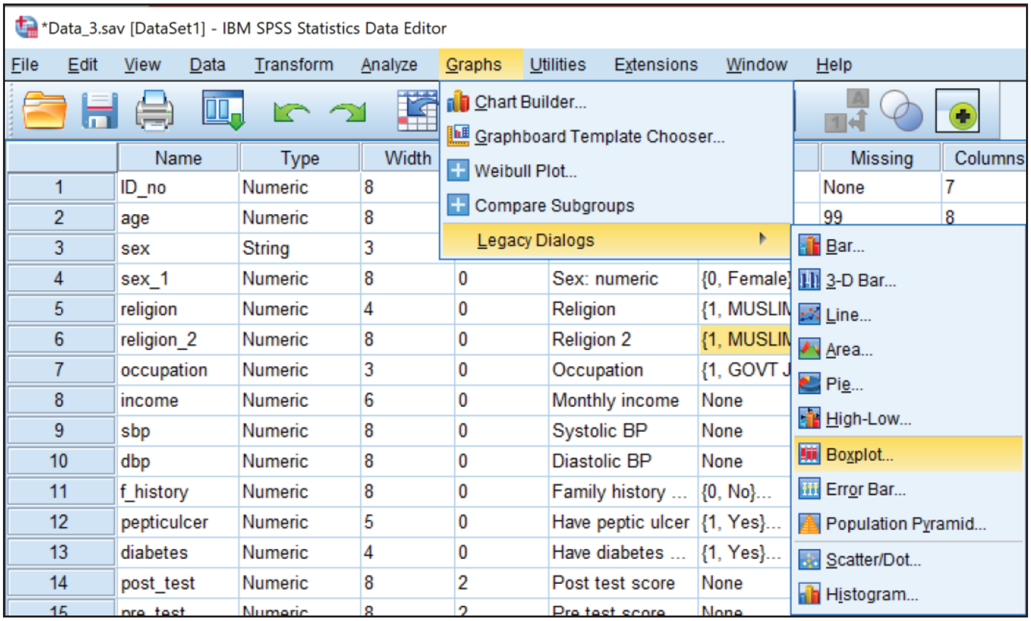
9.1.4 Graph on distribution of medians/means

You can generate the box and plot chart to see the distribution of medians/means of the dependent variable across the groups (at each level of the independent variable, i.e., in different religious groups). To have the box and plot chart, use the following commands.

> Graphs > Legacy dialogs > Boxplot… > Simple > Select "Summaries for groups of cases (already selected by default)" > Define > Select "income" for "Variable" box and select "religion_2" for "Category axis' box > OK (Fig 9.3)

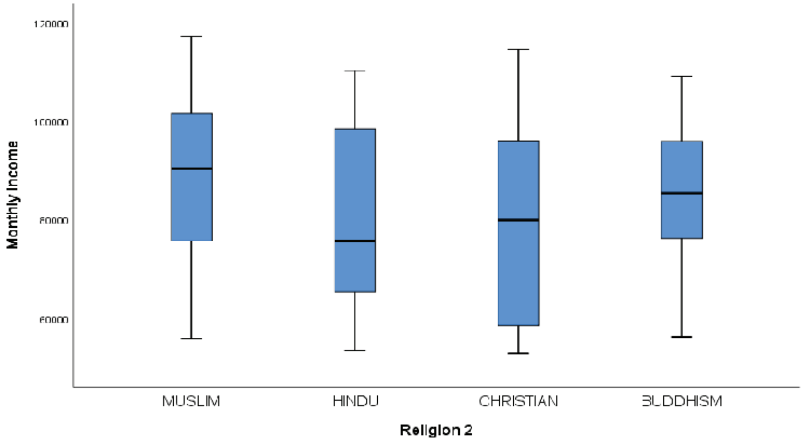
The SPSS will generate the box and plot chart (Fig 9.4) of income by religion. The horizontal line within the box indicates the median.

Figure 9.3



	Name	Type	Width				Missing	Columns
1	ID_no	Numeric	8				None	7
2	age	Numeric	8				99	8
3	sex	String	3					
4	sex_1	Numeric	8	0	Sex: numeric	{0, Female}		
5	religion	Numeric	4	0	Religion	{1, MUSLIM		
6	religion_2	Numeric	8	0	Religion 2	{1, MUSLIM		
7	occupation	Numeric	3	0	Occupation	{1, GOVT J		
8	income	Numeric	6	0	Monthly income	None		
9	sbp	Numeric	8	0	Systolic BP	None		
10	dbp	Numeric	8	0	Diastolic BP	None		
11	f_history	Numeric	8	0	Family history ...	{0, No}...		
12	pepticulcer	Numeric	5	0	Have peptic ulcer	{1, Yes}...		
13	diabetes	Numeric	4	0	Have diabetes ...	{1, Yes}...		
14	post_test	Numeric	8	2	Post test score	None		
15	pre_test	Numeric	8	2	Pre test score	None		

Figure 9.4 Box and plot chart of income by religious groups

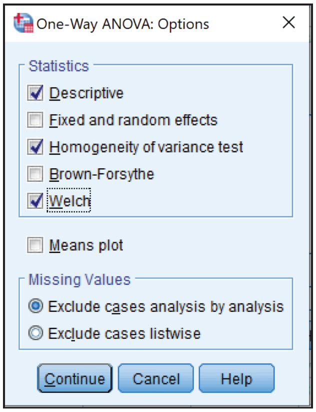


9.1.5 What to do if the variances are not homogeneous?

When the group variances are not homogeneous (i.e., Levene’s test p-value is <0.05) for the comparison of group means, we need to use the *Welch test (or Browne-Forsythe test)*. Similarly, for the comparison of individual group means (post hoc test), instead of Bonferroni’s (or Tukey) test use the *Games-Howell test*. Use the following commands to get these test results.

> Analyze > Compare means > One-way ANOVA > Select “income” and push it into the “Dependent list” box > Select “religion_2” for the “Factor” box > Options > Select “Descriptive”, “Homogeneity of variance test” and “Welch” (Fig 9.5) > Continue > Post Hoc > Select “Games-Howell” under the “Equal Variances not Assumed” (Fig 9.6) > Continue > OK

Figure 9.5



One-Way ANOVA: Options

Statistics

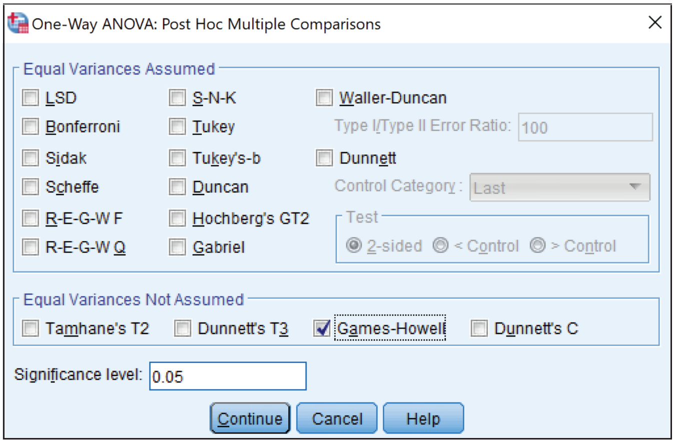
☑ Descriptive
☐ Fixed and random effects
☑ Homogeneity of variance test
☐ Brown-Forsythe
☑ Welch

☐ Means plot

Missing Values

● Exclude cases analysis by analysis
○ Exclude cases listwise

Figure 9.6



One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

☐ LSD ☐ S-N-K ☐ Waller-Duncan
☐ Bonferroni ☐ Tukey Type I/Type II Error Ratio: 100
☐ Sidak ☐ Tukey's-b ☐ Dunnett
☐ Scheffe ☐ Duncan Control Category: Last
☐ R-E-G-W F ☐ Hochberg's GT2 Test
☐ R-E-G-W Q ☐ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

☐ Tamhane's T2 ☐ Dunnett's T3 ☑ Games-Howell ☐ Dunnett's C

Significance level: 0.05

9.1.5.1 Outputs

The SPSS will provide the Welch test results (Table 9.5) and multiple comparisons table (Table 9.6) along with the tables already discussed earlier.

Table 9.5 Welch test for equality of means

Robust Tests of Equality of Means				
Monthly income				
	Statistic[a]	df1	df2	Sig.
Welch	3.292	3	56.236	.027

a. Asymptotically F distributed.

Table 9.6 Comparison of means of income between the religious groups

Multiple Comparisons						
Dependent Variable: Monthly income						
Games-Howell						
		Mean Difference			95% Confidence Interval	
(I) Religion 2	(J) Religion 2	(I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
MUSLIM	HINDU	9014.877*	3340.016	.044	166.37	17863.38
	CHRISTIAN	8775.289	4185.146	.175	-2541.95	20092.53
	BUDDHISM	3384.314	3440.575	.760	-5931.90	12700.53
HINDU	MUSLIM	-9014.877*	3340.016	.044	-17863.38	-166.37
	CHRISTIAN	-239.588	4896.032	1.000	-13248.23	12769.05
	BUDDHISM	-5630.563	4277.059	.557	-16986.14	5725.02
CHRISTIAN	MUSLIM	-8775.289	4185.146	.175	-20092.53	2541.95
	HINDU	239.588	4896.032	1.000	-12769.05	13248.23
	BUDDHISM	-5390.976	4965.176	.700	-18635.83	7853.88
BUDDHISM	MUSLIM	-3384.314	3440.575	.760	-12700.53	5931.90
	HINDU	5630.563	4277.059	.557	-5725.02	16986.14
	CHRISTIAN	5390.976	4965.176	.700	-7853.88	18635.83

*. The mean difference is significant at the 0.05 level.

9.1.5.2 Interpretation

Table 9.5 shows the Welch test results of comparison of means of income in different groups (Robust Tests of Equality of Means). Consider the p-value (Sig.) provided in Table 9.5. The p-value is 0.027, which is <0.05 (reject the null hypothesis). This means that the mean income of all the religious groups is not same in the population.

Table 9.6 conveys the same information as Bonferroni's test that we have discussed earlier. Here, the difference in mean income between Muslims and Hindus is significantly different as indicated by the p-value (Sig.) (p=0.044). The difference of means among the other religious groups is not significant. The table has also provided the 95% CI of the differences.

9.2 Two-way ANOVA

Two-way ANOVA is like one-way ANOVA except that it examines an additional independent categorical variable. Therefore, the two-way ANOVA involves three variables – one quantitative (dependent variable) and two categorical variables. This test is not commonly used in health research. Use the data file <**Data_3.sav**> for practice.

For instance, we want to compare the mean systolic BP (SPSS variable name is "sbp") in different occupation and sex (male and female) groups. Here, the dependent variable is *systolic BP* and the independent categorical variables are *occupation and sex*.
Since we have 4 levels/categories in occupation (govt. job; private job; business and others) and two categories in sex (male and female), we have a factorial design with 8 (4×2) data cells. The two-way ANOVA test answers the following 3 questions:

1. Does occupation influence the systolic BP (i.e., is mean systolic BP among the occupation groups same)?
2. Does sex influence the systolic BP (i.e., is the mean systolic BP same for males and females)?
3. Does the influence of occupation on systolic BP depends on sex (i.e., is there interaction between occupation and sex)?

Questions 1 and 2 refer to the *main effect*, while question 3 explains the *interaction* of two independent variables (occupation and sex) on the dependent variable (systolic BP).

Assumptions

1. The dependent variable (systolic BP) is normally distributed at each level of the independent variables (occupation and sex);
2. The variances of the dependent variable (systolic BP) at each level of the

independent variables are same (homogeneity of variances); and

3. Subjects represent random samples from the populations.

First of all, we need to check the normality of data (systolic BP) in different categories of occupation and sex separately using histogram, Q-Q plot and Shapiro Wilk test (or, K-S test). We must also check the homogeneity of variances in each group of the independent variables (occupation and sex) using the Levene's test.

9.2.1 Commands

Analyze > General linear model > Univariate > Select “sbp” for "Dependent variable" box and select “occupation” and “sex” for “Fixed factors” box > Options > Select “Descriptive statistics, Estimates of effect size and Homogeneity test" > Continue > OK (Fig 9.7 and 9.8)

Figure 9.7

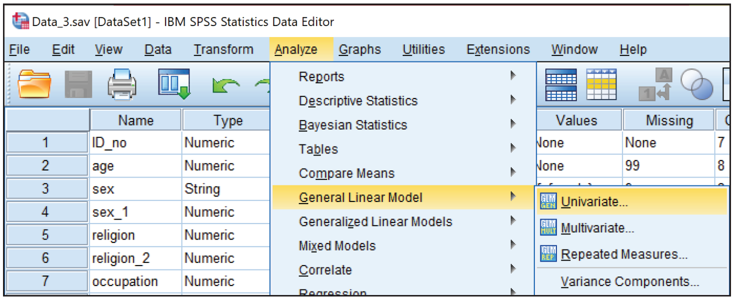
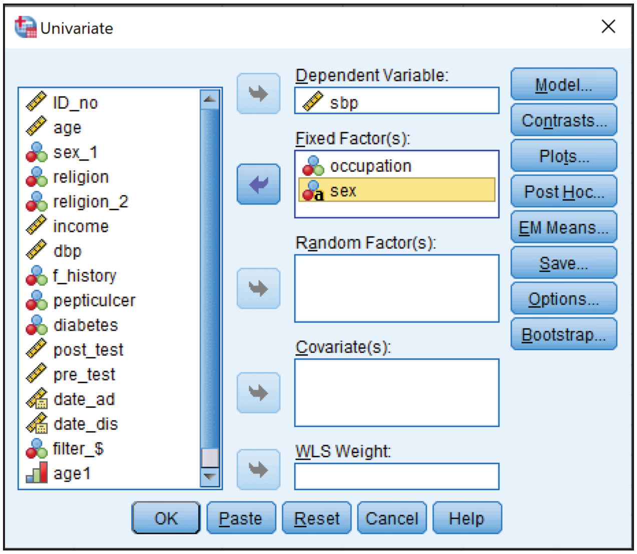


Figure 9.8


Univariate

Dependent Variable: sbp

Fixed Factor(s): occupation, sex

Random Factor(s):

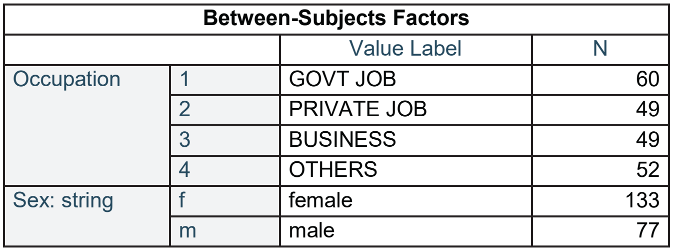
Covariate(s):

WLS Weight:

9.2.2 Outputs

The SPSS will give you the following outputs (Tables 9.7 to 9.10).

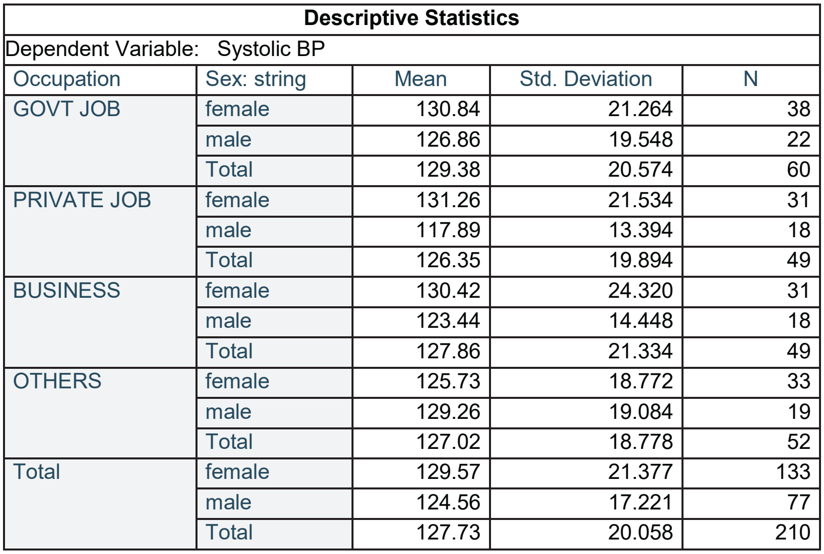
Table 9.7 Frequency distribution of independent variables



Between-Subjects Factors

		Value Label	N
Occupation	1	GOVT JOB	60
	2	PRIVATE JOB	49
	3	BUSINESS	49
	4	OTHERS	52
Sex: string	f	female	133
	m	male	77

Table 9.8 Descriptive statistics of systolic BP by occupation and sex

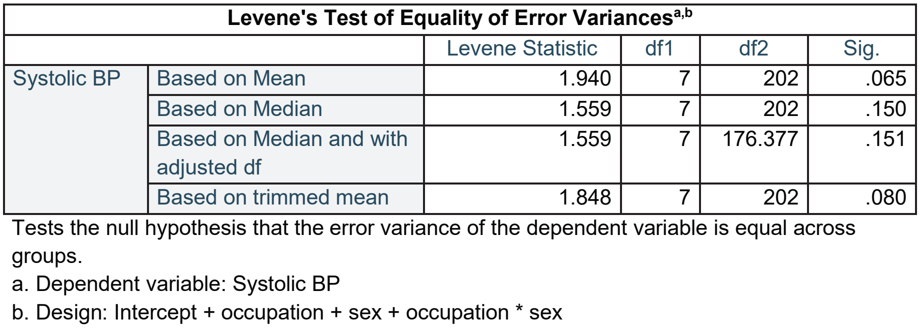


Descriptive Statistics

Dependent Variable: Systolic BP

Occupation	Sex: string	Mean	Std. Deviation	N
GOVT JOB	female	130.84	21.264	38
	male	126.86	19.548	22
	Total	129.38	20.574	60
PRIVATE JOB	female	131.26	21.534	31
	male	117.89	13.394	18
	Total	126.35	19.894	49
BUSINESS	female	130.42	24.320	31
	male	123.44	14.448	18
	Total	127.86	21.334	49
OTHERS	female	125.73	18.772	33
	male	129.26	19.084	19
	Total	127.02	18.778	52
Total	female	129.57	21.377	133
	male	124.56	17.221	77
	Total	127.73	20.058	210

Table 9.9 Levene's test results for equality of variances



Levene's Test of Equality of Error Variances[a,b]

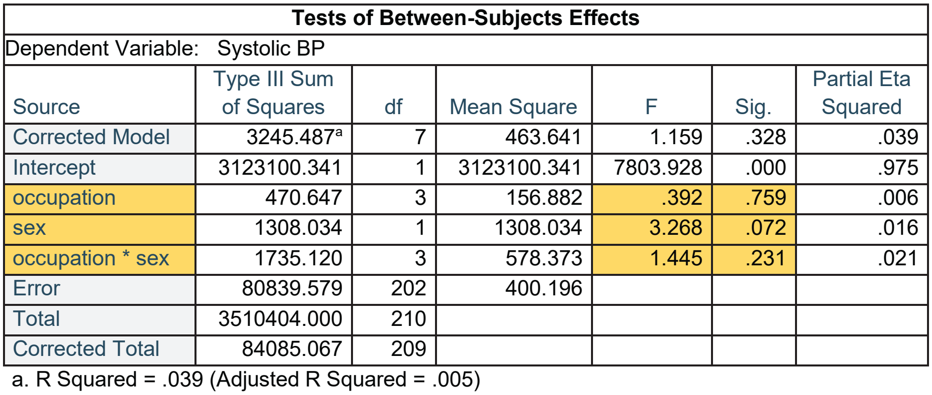
		Levene Statistic	df1	df2	Sig.
Systolic BP	Based on Mean	1.940	7	202	.065
	Based on Median	1.559	7	202	.150
	Based on Median and with adjusted df	1.559	7	176.377	.151
	Based on trimmed mean	1.848	7	202	.080

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: Systolic BP

b. Design: Intercept + occupation + sex + occupation * sex

Table 9.10 The two-way AVOVA table



Tests of Between-Subjects Effects

Dependent Variable: Systolic BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3245.487^a	7	463.641	1.159	.328	.039
Intercept	3123100.341	1	3123100.341	7803.928	.000	.975
occupation	470.647	3	156.882	.392	.759	.006
sex	1308.034	1	1308.034	3.268	.072	.016
occupation * sex	1735.120	3	578.373	1.445	.231	.021
Error	80839.579	202	400.196			
Total	3510404.000	210				
Corrected Total	84085.067	209				

a. R Squared = .039 (Adjusted R Squared = .005)

9.2.3 Interpretation

Table 9.7 (between-subjects factors) shows the frequencies of occupation and sex. Table 9.8 (descriptive statistics) shows the descriptive measures of systolic BP at different levels of occupation and sex. For example, the mean systolic BP of females doing the government job is 130.84 (SD 21.2) mmHg and that of males doing the government job is 126.8 (SD 19.5) mmHg.

Table 9.9 shows the Levene's test results of homogeneity of variances for both the mean and median. Consider the test results for the mean. The p-value (Sig.) of the test, as shown in the table, is 0.065. A p-value >0.05 indicates that the variances of systolic BP at each level of the independent variables (occupation and sex) are not different (homogeneous). Thus, assumption 2 is *not* violated.

The table of "Tests of between-subjects effects" (Table 9.10) shows the *main effects* of the independent variables. Look at the p-values (Sig.) of occupation and sex. They are 0.759 and 0.072, respectively. This indicates that the mean systolic BP is not different in different occupation groups as well as sex (males and females). Now, look at the p-value for *"occupation*sex"*, which indicates the significance of the interaction between these two variables on systolic BP. A p-value ≤0.05 indicates the presence of interaction, that means that the systolic BP of different occupation groups is influenced by (depends on) sex. In our example, the p-value is 0.231 (>0.05), which means that there is no interaction between occupation and sex to influence the systolic BP.

The *Partial Eta Squared* (last column of Table 9.10) indicates the effect size. The Eta statistics for occupation and sex are 0.006 and 0.016, which are very small. These values are equivalent to R^2 (Coefficient of Determination). Eta 0.006 indicates that only 0.6% variance of systolic BP can be explained by occupation (and 1.6% by sex). However, most of the researchers do not report this in their publications.

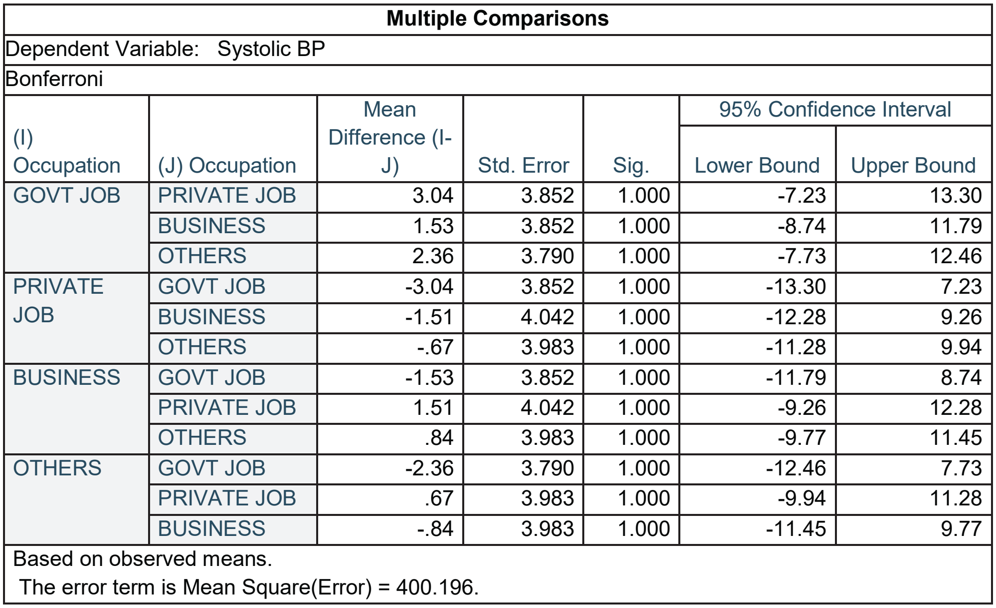
The post-hoc test (as discussed under one-way ANOVA) is performed if the main eff-

-ect is significant (i.e., the p-values for occupation and/or sex are <0.05), otherwise it is not necessary. However, to do the post-hoc test, use the following commands.

> Analyze > General linear model > Univariate > Select “sbp” for "Dependent variable" box and select “occupation” and “sex” for “Fixed factors” box > Options > Select “Descriptive statistics, and Homogeneity test" > Continue > Post hoc > Select “occupation” and “sex” and push them into the “Post Hoc Tests for” box > Select “Bonferroni” under “Equal Variances Assumed” > Continue > OK

This will provide the multiple comparison table of occupation (Table 9.11) along with the other tables as shown earlier. Look at the p-values (Sig.). All the p-values of the differences are 1.00. This means that there is no difference in mean systolic BP between the occupation groups. Note that SPSS did not provide the multiple comparison table for sex, since it has two levels.

Table 9.11 Multiple Comparison table



Multiple Comparisons

Dependent Variable: Systolic BP

Bonferroni

(I) Occupation	(J) Occupation	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
GOVT JOB	PRIVATE JOB	3.04	3.852	1.000	-7.23	13.30
	BUSINESS	1.53	3.852	1.000	-8.74	11.79
	OTHERS	2.36	3.790	1.000	-7.73	12.46
PRIVATE JOB	GOVT JOB	-3.04	3.852	1.000	-13.30	7.23
	BUSINESS	-1.51	4.042	1.000	-12.28	9.26
	OTHERS	-.67	3.983	1.000	-11.28	9.94
BUSINESS	GOVT JOB	-1.53	3.852	1.000	-11.79	8.74
	PRIVATE JOB	1.51	4.042	1.000	-9.26	12.28
	OTHERS	.84	3.983	1.000	-9.77	11.45
OTHERS	GOVT JOB	-2.36	3.790	1.000	-12.46	7.73
	PRIVATE JOB	.67	3.983	1.000	-9.94	11.28
	BUSINESS	-.84	3.983	1.000	-11.45	9.77

Based on observed means.
The error term is Mean Square(Error) = 400.196.

To have a clearer picture of the presence of interaction, it is better to get a graph of the mean systolic BP for occupation and sex. Use the following commands to get the graph.

> Analyze > General linear model > Univariate > Select “sbp” for "Dependent variable" box and select “occupation” and “sex” for “Fixed factors” box > Options > Select “Descriptive statistics, and Homogeneity test" > Continue > Plots > Select “occupation” and push it into the “Horizontal axis” box and select “sex” and push it into the

“Separate lines” box > Add > Select “Line chart” under “Chart type” > Select “Include error bars” and “Confidence interval (95.0%)” under “Error bars” > Continue > OK (Fig 9.9)

You can also draw the line graph using the “Graph” menu as below:

Graphs > Legacy dialogs > Line > Select "Multiple" > Select "Summarizes for groups of cases" > Define > Select "Other statistic (e.g., mean)" > Move the dependent variable (sbp) into the "Variable" box > Select “occupation” for the "Category axis" box (note: select the variable with most categories, here occupation has more categories than sex) > Move “sex” (the other independent variable) into "Define line by" box > OK

The above commands will produce the line graphs of mean systolic BP of different occupation groups as shown in Figures 9.9 and 9.10. The graphs show that there is a greater difference in mean systolic BP between males (117.89 mmHg) and females (131.26 mmHg) among private job holders, compared to other occupations. However, this difference is not statistically significant to show an interaction between occupation and sex. This means that there is no significant variation of systolic BP in the occupation groups as well as there is no influence of sex on occupation groups for systolic BP.

Figure 9.9 Mean systolic BP with 95% CI of different occupation groups by sex

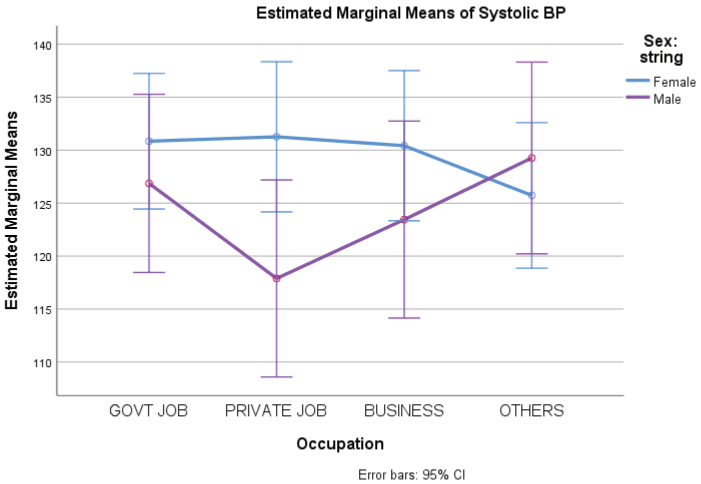
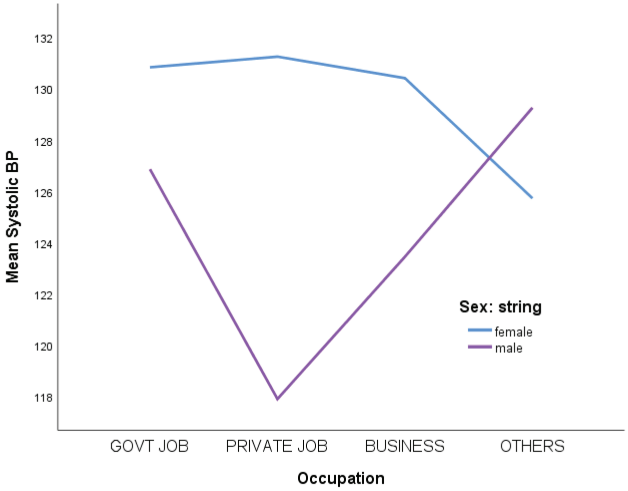


Figure 9.10 Mean systolic BP in different occupation groups by sex



10
Repeated Measures ANOVA

Repeated measures design is a commonly used experimental design in health research. In repeated measures design, measurements of the same variable are made on each subject on two or more different occasions (either at different points in time or under different conditions, such as different treatments). It is similar to a paired t-test, except that there are more than two measurements in repeated measures ANOVA. In this chapter, one-way (within-subjects) and two-way (within and between-subjects) repeated measures ANOVA are discussed.

10.1 Repeated measures ANOVA: One-way

The one-way repeated measures ANOVA test is analogous to paired samples t-test that we have discussed earlier (section 8.3). The main difference is that, in paired samples t-test we have two measurements at different times (e.g., before and after giving a drug, or pre-test and post-test results) on the same subjects, while in one-way repeated measures ANOVA, there are three or more measurements on the same subjects at different points in time (i.e., the subjects are exposed to multiple measurements over a period of time or conditions). The one-way repeated measures ANOVA is also called one-way *within-subjects ANOVA*. Use the data file <**Data_repeat_anova_2.sav**> for practice.

Suppose we are interested to assess the mean blood sugar levels at 4 different time intervals (at hour-0, hour-7, hour-14 and hour-24) after administration of a drug on 15 study subjects. The objective of this study is to assess whether the drug reduces the blood sugar levels over time (i.e., the mean blood sugar levels over time are same or different).

To conduct this study, we have selected 15 individuals randomly from a population and measured their blood sugar levels at the baseline (hour-0). All the individuals are then provided with the drug (say, drug A) and their blood sugar levels are measured again at hour-7, hour-14 and hour-24. We are interested to know if the blood sugar levels over time, after giving the drug, are the same or not (in other words, whether the drug is effec-

-tive in reducing the blood sugar levels over time). The variables, blood sugar levels at hour-0, hour-7, hour-14 and hour-24, are named in SPSS as sugar_0, sugar_7, sugar_14 and sugar_24, respectively. *Note that, in this example, we have only one treatment group (received drug A), but have the outcome measurements (blood sugar) at 4 different points in time on the same subjects (i.e., we have one treatment group with 4 levels of measurements on the same subjects).*

Hypothesis

H_0: The mean blood sugar level is same/equal at each level of measurement (i.e., the population mean of blood sugar at 0, 7, 14 and 24 hours is same).

H_A: The mean blood sugar is not same at different levels of measurement (i.e., population mean of blood sugar at 0, 7, 14 and 24 hours is different).

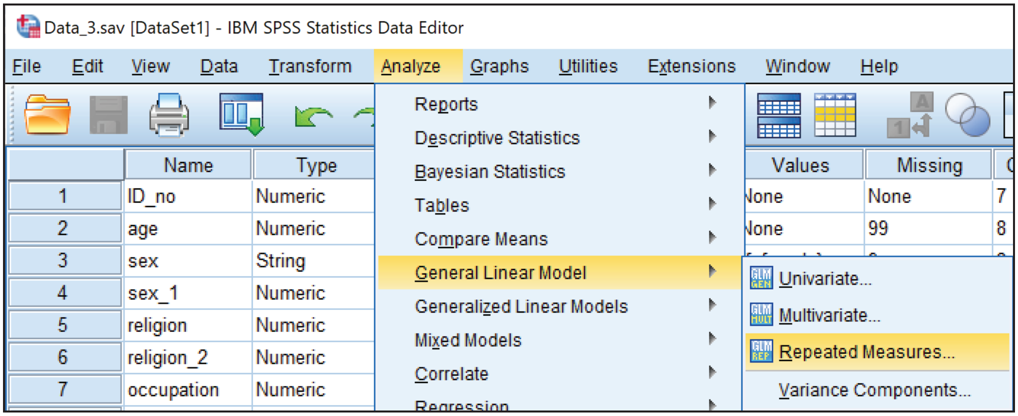
Assumptions

1. The dependent variable (blood sugar level) is normally distributed in the population at each level of within-subjects factor;
2. The population variances of the differences between all combinations of related groups/levels are equal (called *Sphericity assumption*); and
3. The subjects represent a random sample from the population.

10.1.1 Commands

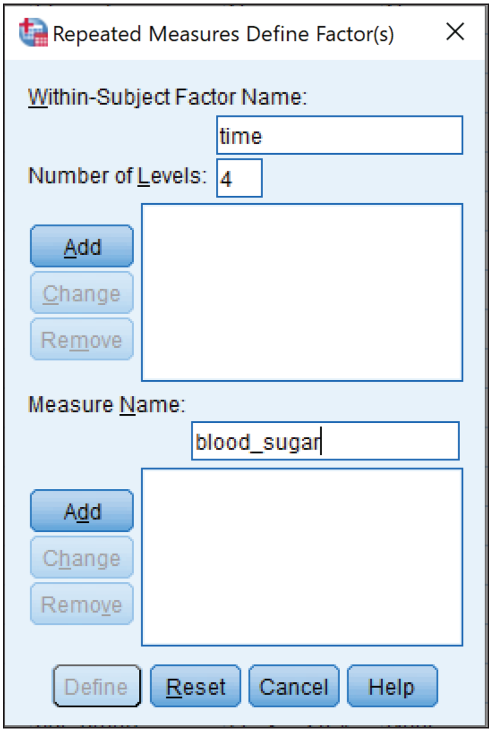
Analyze > General linear model > Repeated measures (Fig 10.1.1) > In “Within subject factor name” box write “time” (or, give any other name) after deleting factor1 > in “Number of levels” box write “4” (since there are 4 time factors) > Add > Write “blood_sugar” in “Measure Name” box (Fig 10.1.2) > Add > Define > Select variables “sugar_0, sugar_7, sugar_14 and sugar_24” and push them into "Within-Subjects Variables (time)" box (Fig 10.1.3) > Options > Select "Descriptive statistics, Estimates of effect size and Homogeneity tests" > Continue > EM Means > Select “time” and push it into the “Display means for” box > Select “Compare main effects” > Select “Bonferroni” in box “Confidence interval adjustment” (Fig 10.1.4) > Continue > Pots > Select “time” and push it into “Horizontal axis” box > Add > Continue > Contrasts > Select “Repeated” from the box “Contrast” > Change > Continue > OK

Figure 10.1.1



	Name	Type	Values	Missing	C
1	ID_no	Numeric	None	None	7
2	age	Numeric	None	99	8
3	sex	String			
4	sex_1	Numeric			
5	religion	Numeric			
6	religion_2	Numeric			
7	occupation	Numeric			

Figure 10.1.2

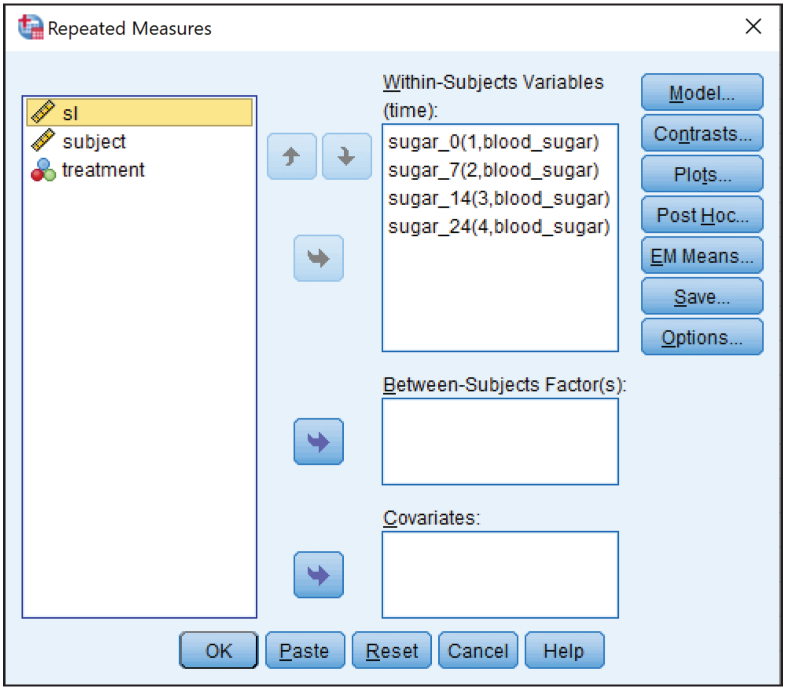


Within-Subject Factor Name: time

Number of Levels: 4

Measure Name: blood_sugar

Figure 10.1.3



Repeated Measures

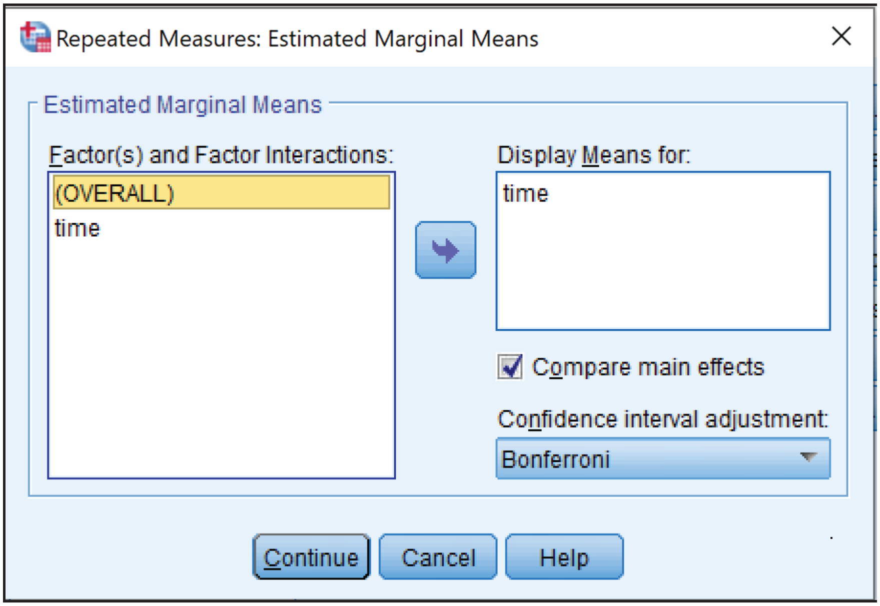
sl
subject
treatment

Within-Subjects Variables (time):
sugar_0(1,blood_sugar)
sugar_7(2,blood_sugar)
sugar_14(3,blood_sugar)
sugar_24(4,blood_sugar)

Between-Subjects Factor(s):

Covariates:

Figure 10.1.4



Repeated Measures: Estimated Marginal Means

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)
time

Display Means for:
time

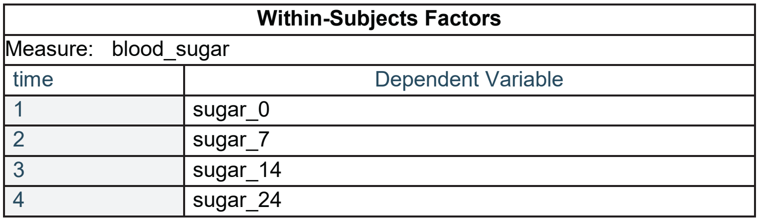
☑ Compare main effects

Confidence interval adjustment:
Bonferroni

10.1.2 Outputs

The SPSS will produce several tables. We need only the following tables (Tables 10.1.1 to 10.1.8) for interpreting the results. The tables are set chronologically for easier interpretation (not in the order as provided by SPSS).

Table 10.1.1 Codes for different levels of measurements of blood sugar

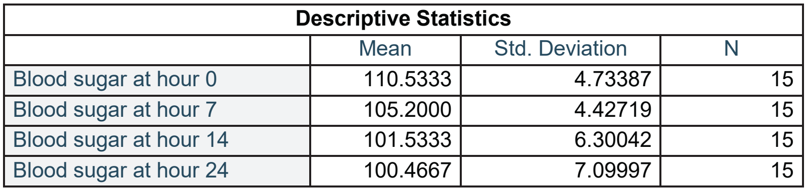


Within-Subjects Factors

Measure: blood_sugar

time	Dependent Variable
1	sugar_0
2	sugar_7
3	sugar_14
4	sugar_24

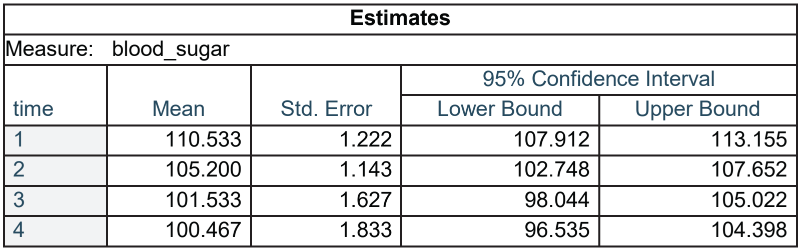
Table 10.1.2 Descriptive statistics of blood sugar at different levels (times) of measurement



Descriptive Statistics

	Mean	Std. Deviation	N
Blood sugar at hour 0	110.5333	4.73387	15
Blood sugar at hour 7	105.2000	4.42719	15
Blood sugar at hour 14	101.5333	6.30042	15
Blood sugar at hour 24	100.4667	7.09997	15

Table 10.1.3 Descriptive statistics of blood sugar at different levels (times) of measurement with 95% CI

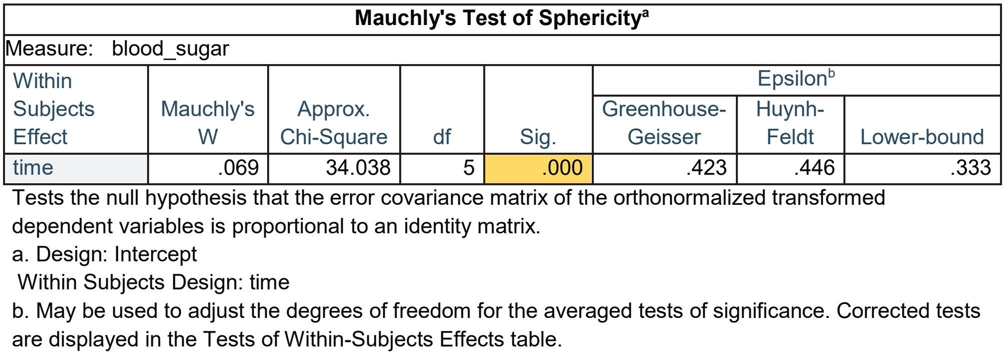


Estimates

Measure: blood_sugar

			95% Confidence Interval	
time	Mean	Std. Error	Lower Bound	Upper Bound
1	110.533	1.222	107.912	113.155
2	105.200	1.143	102.748	107.652
3	101.533	1.627	98.044	105.022
4	100.467	1.833	96.535	104.398

Table 10.1.4 Mauchly's test for Sphericity assumption



Mauchly's Test of Sphericity[a]

Measure: blood_sugar

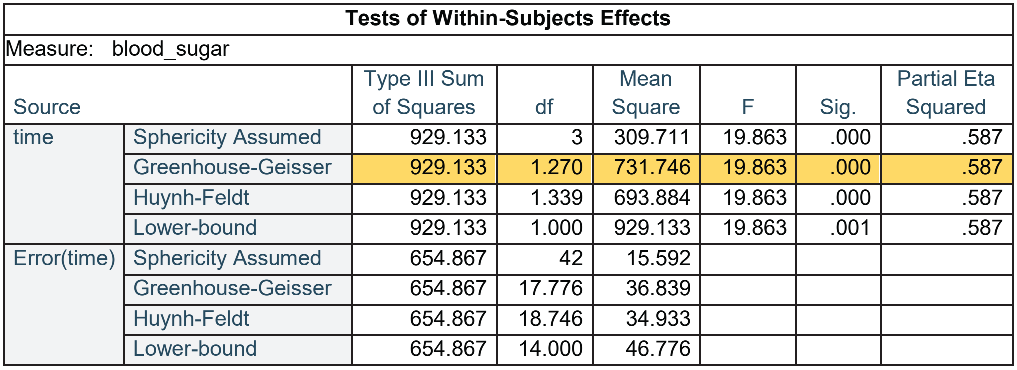
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
time	.069	34.038	5	.000	.423	.446	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: time

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Table 10.1.5 Test results of within subject effects (standard and alternative univariate tests)

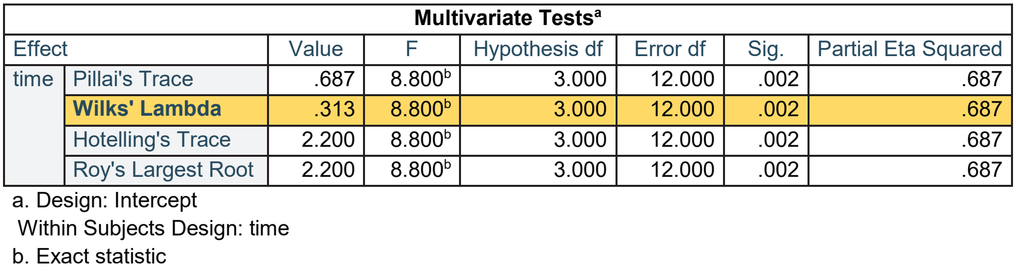


Tests of Within-Subjects Effects

Measure: blood_sugar

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
time	Sphericity Assumed	929.133	3	309.711	19.863	.000	.587
	Greenhouse-Geisser	929.133	1.270	731.746	19.863	.000	.587
	Huynh-Feldt	929.133	1.339	693.884	19.863	.000	.587
	Lower-bound	929.133	1.000	929.133	19.863	.001	.587
Error(time)	Sphericity Assumed	654.867	42	15.592			
	Greenhouse-Geisser	654.867	17.776	36.839			
	Huynh-Feldt	654.867	18.746	34.933			
	Lower-bound	654.867	14.000	46.776			

Table 10.1.6 Multivariate test results



Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
time	Pillai's Trace	.687	8.800[b]	3.000	12.000	.002	.687
	Wilks' Lambda	.313	8.800[b]	3.000	12.000	.002	.687
	Hotelling's Trace	2.200	8.800[b]	3.000	12.000	.002	.687
	Roy's Largest Root	2.200	8.800[b]	3.000	12.000	.002	.687

a. Design: Intercept
Within Subjects Design: time

b. Exact statistic

Table 10.1.7 Pairwise comparison of mean blood sugar at adjacent times of measurement

Tests of Within-Subjects Contrasts

Measure: blood_sugar

Source	time	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
time	Level 1 vs. Level 2	426.667	1	426.667	28.535	.000	.671
	Level 2 vs. Level 3	201.667	1	201.667	16.673	.001	.544
	Level 3 vs. Level 4	17.067	1	17.067	1.485	.243	.096
Error(time)	Level 1 vs. Level 2	209.333	14	14.952			
	Level 2 vs. Level 3	169.333	14	12.095			
	Level 3 vs. Level 4	160.933	14	11.495			

Table 10.1.8 Pair-wise comparison of mean blood sugar at different times of measurement

Pairwise Comparisons

Measure: blood_sugar

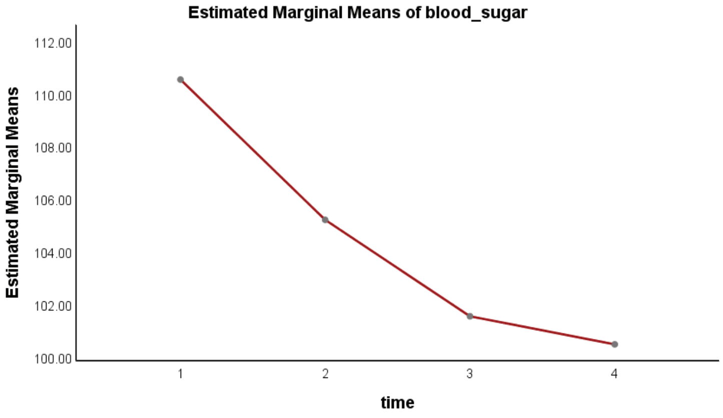
					95% Confidence Interval for Difference[b]	
(I) time	(J) time	Mean Difference (I-J)	Std. Error	Sig.[b]	Lower Bound	Upper Bound
1	2	5.333*	.998	.001	2.269	8.397
	3	9.000*	1.710	.001	3.753	14.247
	4	10.067*	2.216	.003	3.267	16.866
2	1	-5.333*	.998	.001	-8.397	-2.269
	3	3.667*	.898	.007	.911	6.422
	4	4.733*	1.439	.032	.317	9.150
3	1	-9.000*	1.710	.001	-14.247	-3.753
	2	-3.667*	.898	.007	-6.422	-.911
	4	1.067	.875	1.000	-1.620	3.753
4	1	-10.067*	2.216	.003	-16.866	-3.267
	2	-4.733*	1.439	.032	-9.150	-.317
	3	-1.067	.875	1.000	-3.753	1.620

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Figure 10.1.5 Mean blood sugar at different times of measurement



10.1.3 Interpretation

The outputs of the analysis are shown in Tables 10.1.1 to 10.1.8 and Figure 10.1.5. Table 10.1.1 shows the value labels (times) of the blood sugar measurements. Tables 10.1.2 and 10.1.3 (descriptive statistics and estimates) show the descriptive statistics (mean, standard deviation, no. of study subjects, SE of the means, and 95% CI) of the blood sugar levels at different times of measurement, such as at hour-0, hour-7, hour-14 and hour-24 (also displayed in Fig 10.1.5).

One of the important issues for the repeated measures ANOVA test is the Sphericity assumption, as mentioned earlier under "assumptions". Table 10.1.4 shows the test results of "Mauchly's test of Sphericity" to understand whether the Sphericity assumption is correct or violated. The table shows that the Mauchly's W is 0.069 and the p-value is 0.000. Since the p-value is less <0.05, the Sphericity assumption is violated (not correct).

Three types of tests are conducted if within-subjects factors (here, it is the times of measurement of blood sugar, which has 4 levels – hour-0, hour-7, hour-14 and hour-24) have more than 2 levels (here, we have 4 levels). The tests are:

1. Standard univariate test (Sphericity Assumed) [Table 10.1.5];
2. Alternative univariate tests (Greenhouse-Geisser; Huynh-Feldt; Lower-bound) [Table 10.1.5]; and
3. Multivariate tests (Pillai's Trace; Wilks' Lambda; Hotelling's Trace; Roy's Largest Root) [Table 10.1.6]

All these tests evaluate the same hypothesis i.e., the population means are equal at all levels of the measurement. The standard univariate test is based on the *Sphericity assumption,* i.e., the standard univariate test result is considered, if the Sphericity assumption is correct (not violated). *In reality, and in most of the cases (also in our example), the Sphericity assumption is violated, and we cannot use the standard univariate test (Sphericity assumed as given in Table 10.1.5) result*.

In our example, we see that the Sphericity assumption is violated, since the Mauchly's test p-value is 0.000 (Table 10.1.4). Therefore, we shall have to pick up the test results either from alternative univariate tests (Table 10.1.5) or multivariate tests (Table 10.1.6). For practical purposes, it is recommended to use the *multivariate test results* for reporting, since it *does not* depend on the Sphericity assumption.

However, for clarity, let us discuss Table 10.1.5, which indicates the standard and alternative univariate test results. Table 10.1.5 shows the univariate test results of within-subjects effects. The standard univariate ANOVA test result is indicated by the row "Sphericity Assumed". Use this test result, when Sphericity assumption is *correct/not violated* (i.e., Mauchly's test p-value is >0.05). Since our data show that the Sphericity assumption is violated, we cannot use this test result.

When the Sphericity assumption is violated (not correct), you can use the results of one of the alternative univariate tests (i.e., Greenhouse-Geisser, Huynh-Feldt or Lower-bound) for interpretation. It is commonly the Greenhouse-Geisser test, which is reported by the researchers. Table 10.1.5 shows that the test (Greenhouse-Geisser) provided the same F-value and p-value like other tests. Since the test's p-value is 0.000, reject the null hypothesis. This means that the mean blood sugar levels at different time factors (i.e., at different levels of measurement) are not the same.

For simplicity, we would suggest using the *multivariate test results*, which are not dependent on the Sphericity assumption. Table 10.1.6 shows the multivariate test results. In the multivariate tests table, the SPSS has given several test results, such as Pillai's Trace, Wilks' Lambda, Hotelling's Trace and Roy's Largest Root. All these multivariate tests have given the same results. It is recommended to use the *Wilks' Lambda test* results for reporting. In our example, the multivariate test indicates significant time effect on blood sugar levels, as the p-value of Wilks' Lambda is 0.000. This means that the population means of blood sugar levels at different time factors (different times of measurement) are not the same.

Table 10.1.8 shows pairwise comparison of means at different times of measurement. It shows the results as we have seen under one-way ANOVA (Table 9.4; Bonferroni). It is better to consider the differences of adjacent measurements, such as the difference of blood sugar levels between "time 1 and 2", "time 2 and 3" and "time 3 and 4" as shown in Table 10.1.7. The table shows that all the differences have p-values <0.05, except for "time 3 and 4" (p=0.243). This means that mean blood sugar levels are significantly different in all adjacent time periods except for the time between 3 and 4. The mean blood sugar levels at different times of measurement are depicted in Figure 10.1.1.

Note that if the overall test is not significant (i.e., p-value of Wilks' Lambda is >0.05), the table for pairwise comparison is not necessary.

10.2 Repeated measures ANOVA: Two-way

The two-way repeated measures ANOVA is also called *within and between*-subjects ANOVA. In section 10.1, we have discussed the one-way repeated measures ANOVA, which is also called within-subjects ANOVA. In within-subjects ANOVA, we have *only one* treatment (intervention) group. On the other hand, the within and between-subjects ANOVA is used when there is *more than one* treatment group. In this method, at least 3 variables are involved – one dependent *quantitative* variable, and two independent *categorical* variables with two or more levels.

For example, a researcher wants to design an experiment to compare the efficacy of two drugs (to answer which one is more effective) in reducing blood sugar levels over time. In such a situation, the researcher may have the following questions to answer:

1. Is there a difference in mean blood sugar levels between drug A and drug B? This is termed Between-Subjects Factor – a factor that divides the subjects into two or more distinct subgroups.
2. Is there a reduction in mean blood sugar levels over a time period? This is termed Within-Subjects Factor – distinct measurements are made on the same subjects over time. For example, blood sugar levels over time or blood pressure over time, etc.
3. Is there a group-time interaction? If there is a time trend, and whether this trend exists for all groups or only for certain groups?

To answer these questions, we can use *within and between-subjects repeated measures ANOVA*.

Suppose the researcher has decided to compare the efficacy of the drugs Daonil (Glibenclamide) and Metformin (these drugs are used for the treatment of diabetes mellitus) for the reduction of blood sugar levels. In this example, there are two treatment groups (SPSS variable name is "treatment") – Daonil and Metformin. To do the experiment, the researcher has selected 10 subjects and randomly allocated the treatments (5 in each group). The blood sugar levels of the subjects were measured at the baseline (sugar_0), after 7 hours (sugar_7), after 14 hours (sugar_14) and after 24 hours (sugar_24). Data is provided in the data file <**Data_Repeat_anova_2.sav**>.

Hypothesis

We test two hypotheses here. One is for within-subjects effects and the other is for between-subjects effects.

H_0: Daonil and Metformin are equally effective in reducing the blood sugar levels over time (between-subjects effects).

H_A: Both these drugs are not equally effective in reducing the blood sugar levels over time (you can also use one-sided hypothesis, such as "Daonil is more effective in reducing blood sugar levels over time compared to Metformin").

We can also test the hypothesis of whether these drugs are effective in reducing blood sugar levels over time (within-subjects effects; discussed in section 10.1). The assumptions of two-way repeated measures ANOVA are same as those of one-way repeated measures ANOVA.

10.2.1 Commands

Analyze > General linear model > Repeated measures > Write “Time” (or, give any other name) in “Within subject factor name” box after deleting “factor1” > Write “4” in “Number of levels” box (since we have 4 time levels) > Add > Write “Blood_sugar” in “Measures name” box > Add > Define > Select variables “sugar_0, sugar_7, sugar_14 and sugar_24” and push them into "Within-Subjects Variables" box > Select “treatment” and push it into “Between-subjects factors” box > Options > Select "Descriptive statistics” and “Homogeneity tests" > Continue > EM Means > Select “treatment” and “Time” and push them into the “Display means for” box > Select “Compare main effects” > Select “Bonferroni” in “Confidence interval adjustment” box > Continue > Contrasts > Select “time” > Select “Repeated” in the “Contrast” box under “Change contrast” > Change > Continue > Plots > Select “Time” and push it into “Horizontal axis” box > Select “treatment” and push it into the “Separate lines” box > Add > Continue > OK

10.2.2 Outputs

The SPSS will provide many tables, but only the relevant ones are provided below. The outputs are arranged according to – A) Basic tables; B) Tables related to Within-subjects effects; C) Tables related to Between-subjects effects; D) Tables to check the assumptions; and E) Additional tables.

A. Basic tables (Tables 10.2.1 to 10.2.3):

Table 10.2.1 Codes for different levels of measurement

Within-Subjects Factors	
Measure: Blood_sugar	
Time	Dependent Variable
1	sugar_0
2	sugar_7
3	sugar_14
4	sugar_24

Table 10.2.2 Codes of different treatment groups

Between-Subjects Factors			
		Value Label	N
treatment groups	1	Daonil	5
	2	Metformin	5

Table 10.2.3 Descriptive statistics of blood sugar at different levels and treatment groups

Descriptive Statistics				
	treatment groups	Mean	Std. Deviation	N
Blood sugar at hour 0	Daonil	112.8000	2.16795	5
	Metformin	108.4000	7.09225	5
	Total	110.6000	5.46097	10
Blood sugar at hour 7	Daonil	104.0000	4.18330	5
	Metformin	103.0000	4.69042	5
	Total	103.5000	4.22295	10
Blood sugar at hour 14	Daonil	97.4000	3.43511	5
	Metformin	98.6000	3.91152	5
	Total	98.0000	3.52767	10
Blood sugar at hour 24	Daonil	94.4000	2.70185	5
	Metformin	97.6000	2.50998	5
	Total	96.0000	2.98142	10

B. Within-subjects effects (Tables 10.2.4 to 10.2.6):

Table 10.2.4 Within-subjects multivariate test results

Multivariate Tests[a]						
Effect		Value	F	Hypothesis df	Error df	Sig.
Time	Pillai's Trace	.955	42.767[b]	3.000	6.000	.000
	Wilks' Lambda	.045	42.767[b]	3.000	6.000	.000
	Hotelling's Trace	21.384	42.767[b]	3.000	6.000	.000
	Roy's Largest Root	21.384	42.767[b]	3.000	6.000	.000
Time * treatment	Pillai's Trace	.452	1.649[b]	3.000	6.000	.275
	Wilks' Lambda	.548	1.649[b]	3.000	6.000	.275
	Hotelling's Trace	.825	1.649[b]	3.000	6.000	.275
	Roy's Largest Root	.825	1.649[b]	3.000	6.000	.275

a. Design: Intercept + treatment
Within Subjects Design: Time

b. Exact statistic

Table 10.2.5 Descriptive measures of blood sugar at different levels of time with 95% CI

2. Time				
Measure: Blood_sugar				
			95% Confidence Interval	
Time	Mean	Std. Error	Lower Bound	Upper Bound
1	110.600	1.658	106.776	114.424
2	103.500	1.405	100.259	106.741
3	98.000	1.164	95.316	100.684
4	96.000	.825	94.098	97.902

Table 10.2.6 Pairwise comparisons of adjacent blood sugar levels at different time intervals

Tests of Within-Subjects Contrasts

Measure: Blood_sugar

Source	Time	Type III Sum of Squares	df	Mean Square	F	Sig.
Time	Level 1 vs. Level 2	504.100	1	504.100	48.010	.000
	Level 2 vs. Level 3	302.500	1	302.500	99.180	.000
	Level 3 vs. Level 4	40.000	1	40.000	4.000	.081
Time * treatment	Level 1 vs. Level 2	28.900	1	28.900	2.752	.136
	Level 2 vs. Level 3	12.100	1	12.100	3.967	.082
	Level 3 vs. Level 4	10.000	1	10.000	1.000	.347
Error(Time)	Level 1 vs. Level 2	84.000	8	10.500		
	Level 2 vs. Level 3	24.400	8	3.050		
	Level 3 vs. Level 4	80.000	8	10.000		

C. Between-subjects effects (Tables 10.2.7 to 10.2.9 and Fig 10.2.1):

Table 10.2.7 Test results of between-subjects effects

Tests of Between-Subjects Effects

Measure: Blood_sugar

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	104091.006	1	104091.006	9211.593	.000
treatment	.156	1	.156	.014	.909
Error	90.400	8	11.300		

Table 10.2.8 Descriptive statistics of treatment groups

1. treatment groups

Measure: Blood_sugar

treatment groups	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Daonil	102.150	1.503	98.683	105.617
Metformin	101.900	1.503	98.433	105.367

Table 10.2.9 Pairwise comparisons by treatment groups

Pairwise Comparisons

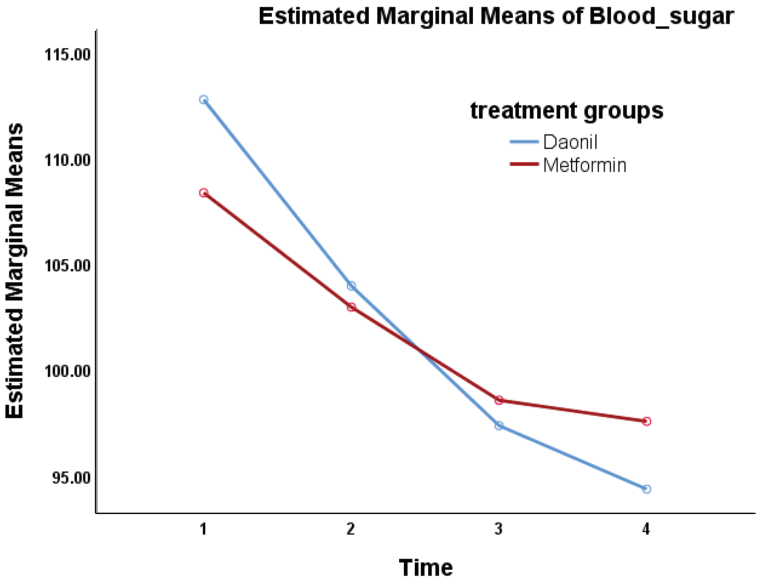
Measure: Blood_sugar

(I) treatment groups	(J) treatment groups	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Daonil	Metformin	.250	2.126	.909	-4.653	5.153
Metformin	Daonil	-.250	2.126	.909	-5.153	4.653

Based on estimated marginal means

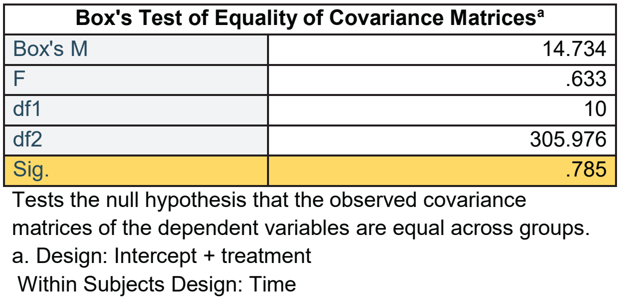
a. Adjustment for multiple comparisons: Bonferroni.

Figure 10.2.1 Blood sugar levels by treatment group (Daonil and Metformin)



D. Tables for checking assumptions (Tables 10.2.10 to 10.2.12):

Table 10.2.10 Box's M test

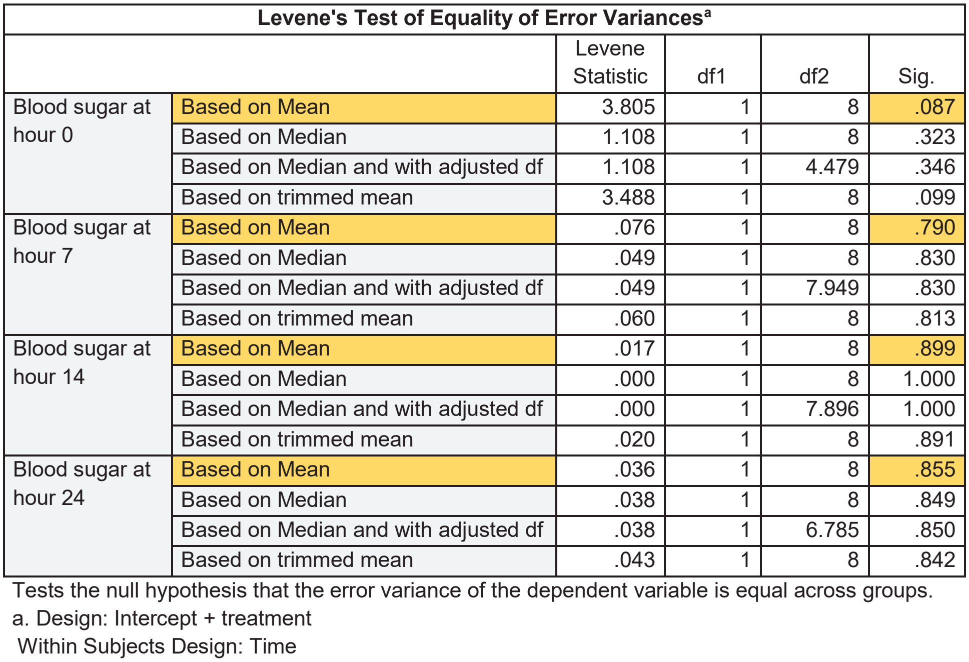


Box's Test of Equality of Covariance Matrices[a]	
Box's M	14.734
F	.633
df1	10
df2	305.976
Sig.	.785

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + treatment
Within Subjects Design: Time

Table 10.2.11 Levene's test of equality of variances



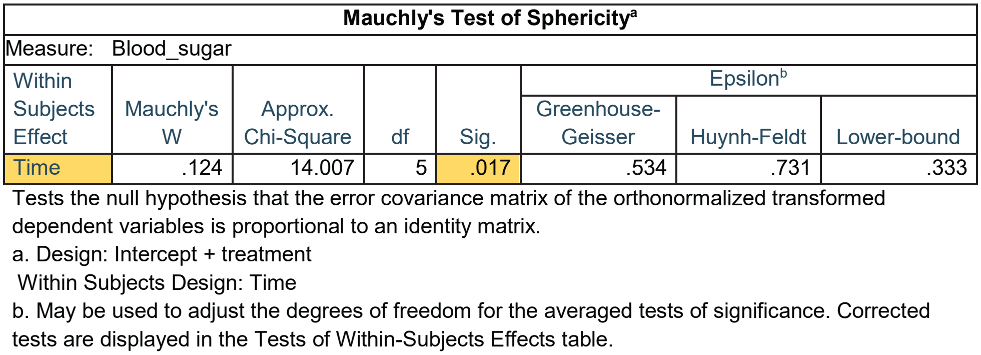
Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
Blood sugar at hour 0	Based on Mean	3.805	1	8	.087
	Based on Median	1.108	1	8	.323
	Based on Median and with adjusted df	1.108	1	4.479	.346
	Based on trimmed mean	3.488	1	8	.099
Blood sugar at hour 7	Based on Mean	.076	1	8	.790
	Based on Median	.049	1	8	.830
	Based on Median and with adjusted df	.049	1	7.949	.830
	Based on trimmed mean	.060	1	8	.813
Blood sugar at hour 14	Based on Mean	.017	1	8	.899
	Based on Median	.000	1	8	1.000
	Based on Median and with adjusted df	.000	1	7.896	1.000
	Based on trimmed mean	.020	1	8	.891
Blood sugar at hour 24	Based on Mean	.036	1	8	.855
	Based on Median	.038	1	8	.849
	Based on Median and with adjusted df	.038	1	6.785	.850
	Based on trimmed mean	.043	1	8	.842

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + treatment
Within Subjects Design: Time

Table 10.2.12 Mauchly's test for Sphericity assumption



Mauchly's Test of Sphericity[a]

Measure: Blood_sugar

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Time	.124	14.007	5	.017	.534	.731	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

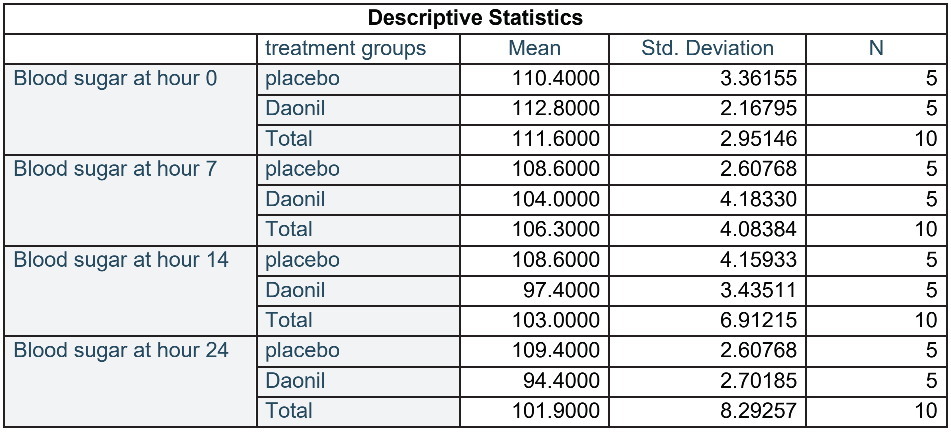
a. Design: Intercept + treatment
Within Subjects Design: Time

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

E. Additional tables and figure: Comparison between Daonil and Placebo

Here, we have analyzed data to compare the effect of Daonil compared to Placebo in reducing blood sugar levels. The following tables (Tables 10.2.13 to 10.2.17) and figure (Fig 10.2.2) show the results of the analysis. In this example, the treatment groups are *significantly different* in reducing blood sugar levels.

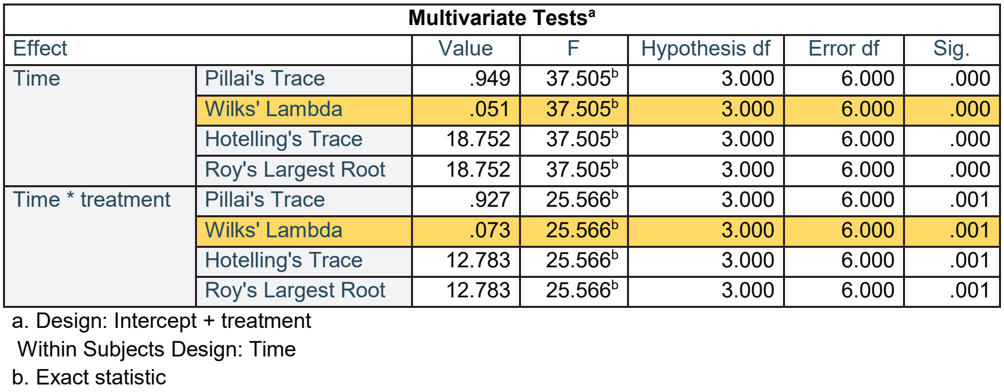
Table 10.2.13 Descriptive statistics of treatment groups (Daonil and Placebo)



Descriptive Statistics

	treatment groups	Mean	Std. Deviation	N
Blood sugar at hour 0	placebo	110.4000	3.36155	5
	Daonil	112.8000	2.16795	5
	Total	111.6000	2.95146	10
Blood sugar at hour 7	placebo	108.6000	2.60768	5
	Daonil	104.0000	4.18330	5
	Total	106.3000	4.08384	10
Blood sugar at hour 14	placebo	108.6000	4.15933	5
	Daonil	97.4000	3.43511	5
	Total	103.0000	6.91215	10
Blood sugar at hour 24	placebo	109.4000	2.60768	5
	Daonil	94.4000	2.70185	5
	Total	101.9000	8.29257	10

Table 10.2.14 Multivariate test results of within-subjects effects



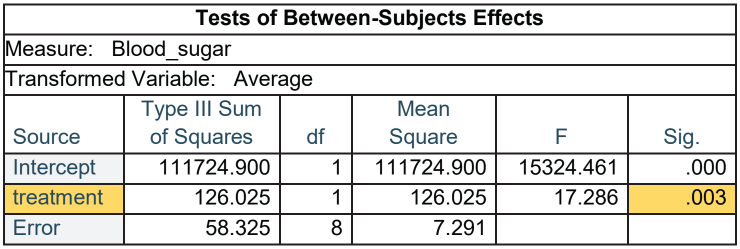
Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Time	Pillai's Trace	.949	37.505[b]	3.000	6.000	.000
	Wilks' Lambda	.051	37.505[b]	3.000	6.000	.000
	Hotelling's Trace	18.752	37.505[b]	3.000	6.000	.000
	Roy's Largest Root	18.752	37.505[b]	3.000	6.000	.000
Time * treatment	Pillai's Trace	.927	25.566[b]	3.000	6.000	.001
	Wilks' Lambda	.073	25.566[b]	3.000	6.000	.001
	Hotelling's Trace	12.783	25.566[b]	3.000	6.000	.001
	Roy's Largest Root	12.783	25.566[b]	3.000	6.000	.001

a. Design: Intercept + treatment
Within Subjects Design: Time

b. Exact statistic

Table 10.2.15 Test results of between-subjects effects



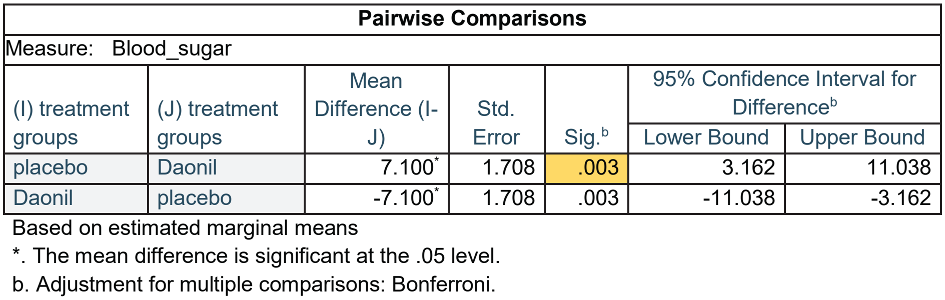
Tests of Between-Subjects Effects

Measure: Blood_sugar

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	111724.900	1	111724.900	15324.461	.000
treatment	126.025	1	126.025	17.286	.003
Error	58.325	8	7.291		

Table 10.2.16 Pairwise comparison of blood sugar levels by treatment groups



Pairwise Comparisons

Measure: Blood_sugar

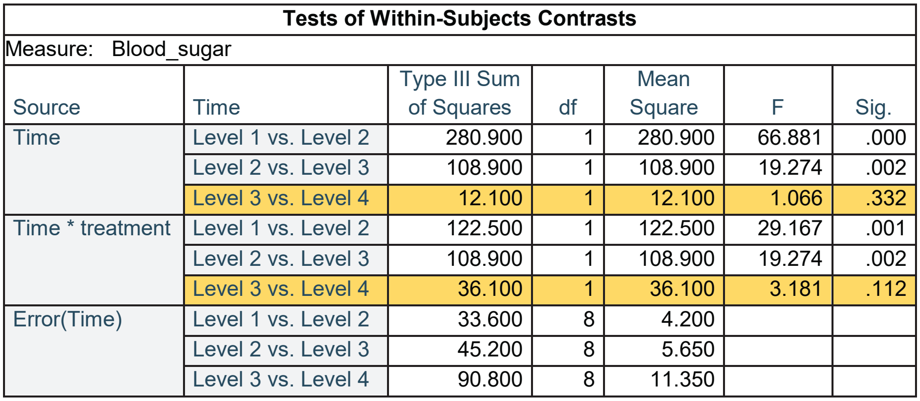
(I) treatment groups	(J) treatment groups	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
placebo	Daonil	7.100*	1.708	.003	3.162	11.038
Daonil	placebo	-7.100*	1.708	.003	-11.038	-3.162

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Table 10.2.17 Pairwise comparison of mean blood sugar levels at adjacent time periods

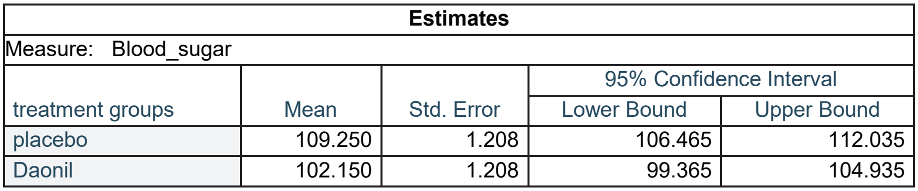


Tests of Within-Subjects Contrasts

Measure: Blood_sugar

Source	Time	Type III Sum of Squares	df	Mean Square	F	Sig.
Time	Level 1 vs. Level 2	280.900	1	280.900	66.881	.000
	Level 2 vs. Level 3	108.900	1	108.900	19.274	.002
	Level 3 vs. Level 4	12.100	1	12.100	1.066	.332
Time * treatment	Level 1 vs. Level 2	122.500	1	122.500	29.167	.001
	Level 2 vs. Level 3	108.900	1	108.900	19.274	.002
	Level 3 vs. Level 4	36.100	1	36.100	3.181	.112
Error(Time)	Level 1 vs. Level 2	33.600	8	4.200		
	Level 2 vs. Level 3	45.200	8	5.650		
	Level 3 vs. Level 4	90.800	8	11.350		

Table 10.2.18 Descriptive statistics by treatment type

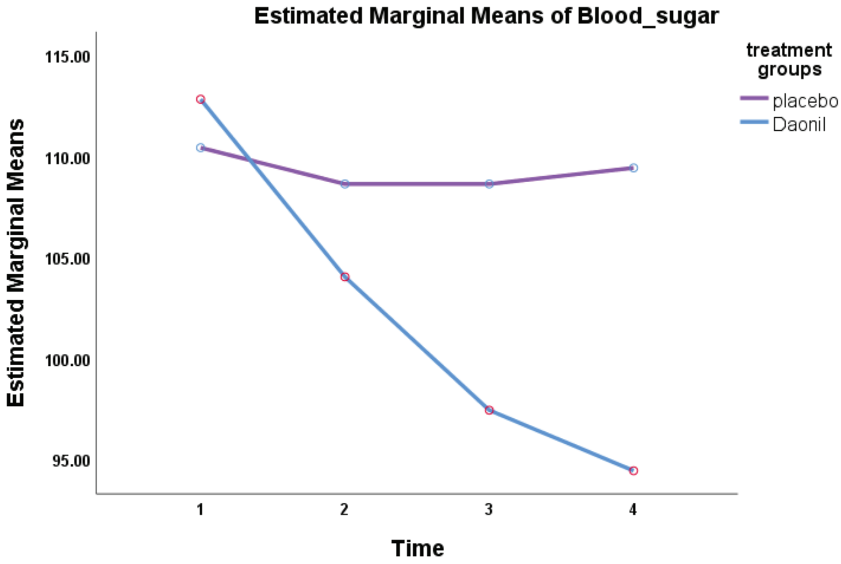


Estimates

Measure: Blood_sugar

treatment groups	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
placebo	109.250	1.208	106.465	112.035
Daonil	102.150	1.208	99.365	104.935

Figure 10.2.2 Blood sugar levels by treatment group (Placebo & Daonil)



10.2.3 Interpretation

A. Basic tables (outputs under A)

The outputs of the analysis are shown in tables and graphs. Table 10.2.1 and 10.2.2 show the value labels of the blood sugar measurements and treatment groups, respectively. Table 10.2.3 shows the descriptive statistics (mean, standard deviation and no. of study subjects) of blood sugar levels at different times of measurement by treatment groups. For example, the mean baseline (hour 0) blood sugar levels were 112.8 (SD 2.1) and 108.4 (SD 7.0) in Daonil and Metformin group, respectively.

B. Within-subjects effects (outputs under B)

Table 10.2.4 shows the multivariate test results of within-subjects effects. Since the Sphericity assumption is frequently violated, we would consider the multivariate test results (Table 10.2.4) as discussed in section 10.1. First, look at the interaction term (Time*treatment) in the row of Wilks' Lambda. The p-value (Sig.) is 0.275, which is not significant. This means that there is no interaction between "Time" and "treatment" (i.e., blood sugar levels over time are not dependent on the treatment groups). Now, look at the row of Wilks' Lambda at "Time". The p-value is 0.000, which is statistically significant. This means that the mean blood sugar levels measured at different times are significantly different (i.e., there is a significant reduction in blood sugar levels over time in both the treatment groups as shown in Fig 10.2.1). Table 10.2.5 shows the means and 95% confid-

-ence intervals of blood sugar levels at different times of measurement. Table 10.2.6 shows the differences in blood sugar levels between the adjacent measurements. The table shows that there are significant differences in blood sugar levels between time 1 and 2 (level 1 vs. level 2), and time 2 and 3 (p=0.000), but not between time 3 and 4 (p=0.081). Note that the interaction (Time*treatment) for any comparison is not significant.

C. Between-subjects effects (outputs under C)

Our primary interest is in the between-subjects effects, since we want to test the hypothesis which treatment is more effective in reducing blood sugar levels over time. Table 10.2.7 shows the results of between-subjects effects (between two treatment groups – Daonil and Metformin) on blood sugar levels. The p-value for treatment is 0.909, which is >0.05. This indicates that there is no significant difference in mean blood sugar levels over time, between Daonil and Metformin groups (also see Fig 10.2.1); i.e., we are unable to reject the null hypothesis. We can see in Table 10.2.8 that the overall means (without considering the times of measurement) of blood sugar levels for Daonil and Metformin are not that different (102.1 vs. 101.9). This indicates that both the drugs are equally effective in reducing the blood sugar levels (no one is better than the other).

The pairwise comparison (Table 10.2.9) shows that the difference in mean blood sugar levels between Daonil and Metformin (0.250) is very small to reject the null hypothesis (p=0.909; Table 10.2.7). Figure 10.2.1 shows the mean blood sugar levels at different times by treatment groups. It shows that blood sugar levels have been reduced by both the drugs over time, but the difference in reduction between the groups is not significant.

D. Test of assumptions (outputs under D)

Whether the assumptions are violated or not are checked by: a) Box's M test (Table 10.2.10); b) Levene's test (Table 10.2.11); and c) Mauchly's test (Table 10.2.12). If the assumptions are met, the p-values of all these tests would be >0.05. We can see that the p-values of all these tests are >0.05, except for Mauchly's test (p=0.017). Note that the Mauchly's test, tests the Sphericity assumption. As discussed earlier, to interpret the results, it is recommended to use the multivariate test results, which are not dependent on Sphericity assumption.

E. Additional tables (outputs under E)

Additional tables (Tables 10.2.13 to 10.2.17) are provided to demonstrate the results, when the treatment groups are different (one is better than the other). Here, we have compared the effectiveness of Daonil compared to Placebo in reducing blood sugar levels over time. Table 10.2.13 shows the descriptive statistics of blood sugar levels at different time intervals by treatment group (Placebo and Daonil). Though there is no significant difference in mean blood sugar levels at hour 0 (baseline) between the treatment groups (110.4 vs. 112.8), but they are different over time (Fig 10.2.2).

The multivariate test results of within-subjects effects (Table 10.2.14) show that there is interaction (time*treatment) between “time” and “treatment” (p=0.001) [see the row of Wilks’ Lambda under Time*treatment]. The p-value of Wilks’ Lambda under “time” is also significant (p=0.000). This means that there is a significant reduction in mean blood sugar levels over time, and it depends on the treatment group (since there is an interaction between time and treatment). Look at Figure 10.2.2. It shows that blood sugar levels have been reduced significantly over time, among the subjects under the treatment of Daonil, but there is no significant change in the placebo group. The test of between-subjects effects (Table 10.2.15) also conveys the information that the difference in blood sugar levels over time is not same for Daonil and placebo groups since the p-value is <0.05 (p-value of the test is 0.003). We, therefore, conclude that Daonil is effective in reducing blood sugar levels and is superior to (better than) Placebo (p=0.003). Table 10.2.16 shows the pairwise comparison of blood sugar levels by treatment groups, while Table 10.2.17 shows the pairwise comparison of mean blood sugar levels at adjacent time periods. If you want to assess the effect size (Partial Eta Squared), select “Effect Size” from the “Option” template during analysis.

11

Association between Two Categorical Variables: Chi-Square Test of Independence

The Chi-square test is a commonly used statistical test for testing hypothesis in health and social sciences research. This test is suitable to determine the association between two categorical variables, whether the data are from cross-sectional, case-control or cohort studies. On the other hand, in epidemiology, cross-tabulations are commonly done to calculate the Odds Ratio (OR) [e.g., for case-control studies] or Relative Risk (RR) [e.g., for cohort studies] with 95% Confidence Intervals (CI). Odds Ratio and RR are the measures of strength of association between two variables. Use the data file <**Data_3.sav**> for practice.

11.1 Chi-square test of independence

Suppose you have collected data on gender (sex) and diabetes from a group of individuals selected randomly from a population. You are interested to know if there is an association between gender and diabetes. In this scenario, Chi-square test is the appropriate test for testing the hypothesis.

Hypothesis

$\mathbf{H_0}$: There is no association between gender and diabetes (it can also be stated as gender and diabetes are independent).

$\mathbf{H_A}$: There is an association between gender and diabetes (or, gender and diabetes are not independent).

Assumption

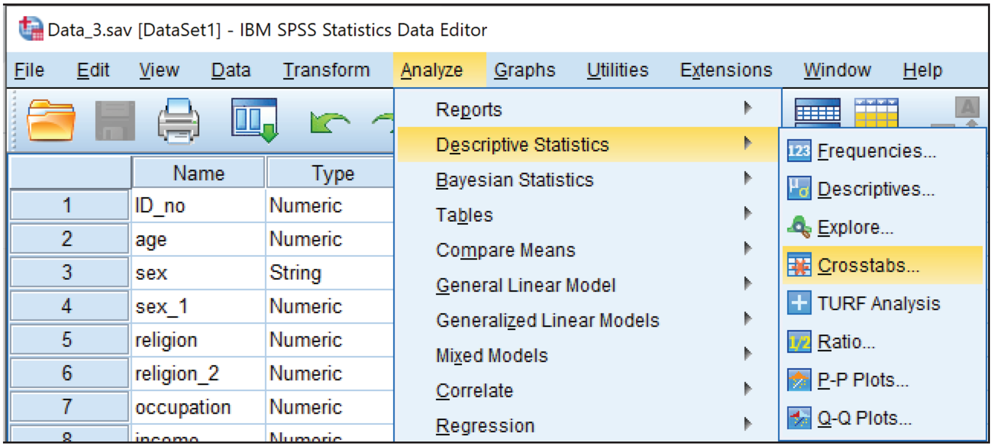
1. Data have come from a random sample drawn from a selected population.

11.1.1 Commands

Analyze > Descriptive statistics > Crosstabs > Select “gender” and push it into the “Row(s)” box > Select “diabetes” for the “Column(s)” box > Statistics > Select “Chi-square” and “Risk” > Continue > Cells > Select “Raw” and “Column” under “Percentages” > Continue > OK (Figs 11.1 to 11.4)

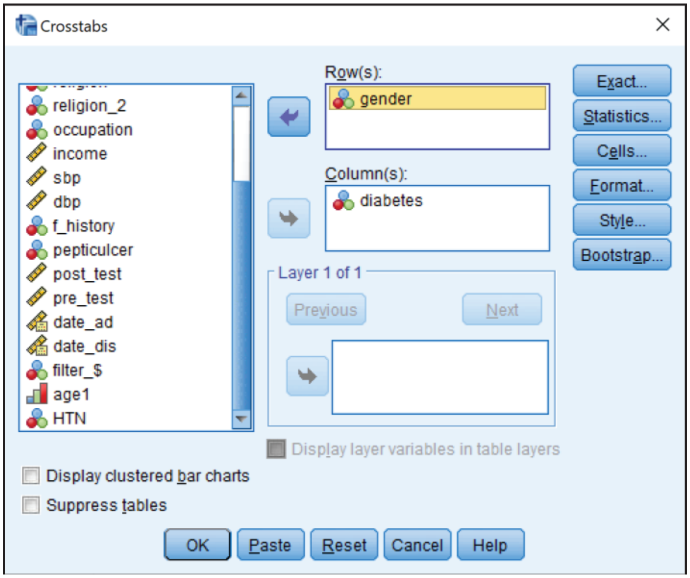
Note: We have selected “risk” to get the OR and RR including their 95% CIs.

Figure 11.1



	Name	Type
1	ID_no	Numeric
2	age	Numeric
3	sex	String
4	sex_1	Numeric
5	religion	Numeric
6	religion_2	Numeric
7	occupation	Numeric
8	income	Numeric

Figure 11.2



Row(s): gender

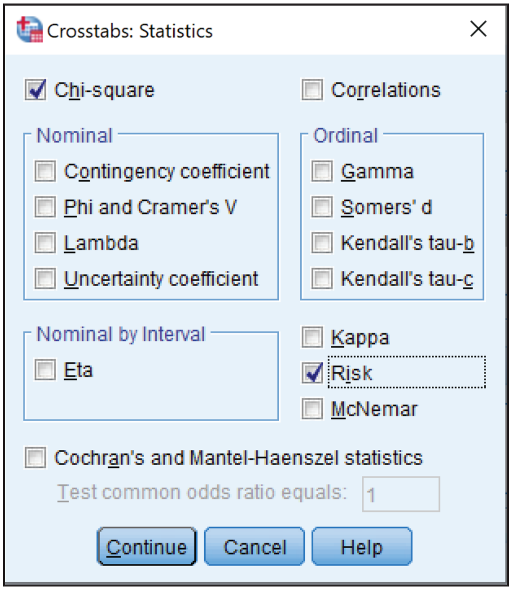
Column(s): diabetes

□ Display layer variables in table layers

□ Display clustered bar charts

□ Suppress tables

Figure 11.3



Crosstabs: Statistics

☑ Chi-square
□ Correlations

Nominal
- □ Contingency coefficient
- □ Phi and Cramer's V
- □ Lambda
- □ Uncertainty coefficient

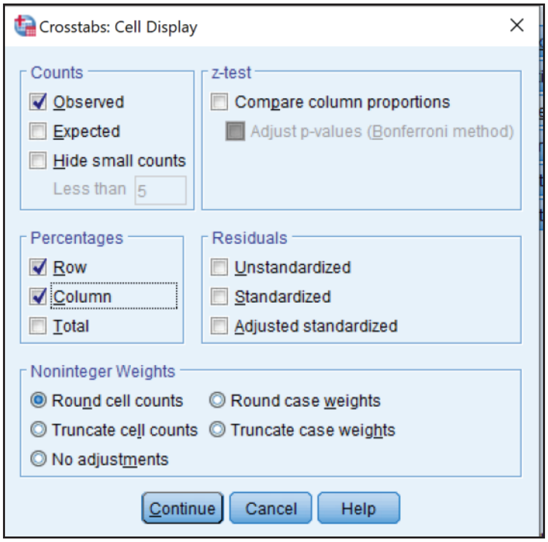
Ordinal
- □ Gamma
- □ Somers' d
- □ Kendall's tau-b
- □ Kendall's tau-c

Nominal by Interval
- □ Eta

□ Kappa
☑ Risk
□ McNemar

□ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1

Figure 11.4



Crosstabs: Cell Display

Counts
- ☑ Observed
- □ Expected
- □ Hide small counts
- Less than 5

z-test
- □ Compare column proportions
- □ Adjust p-values (Bonferroni method)

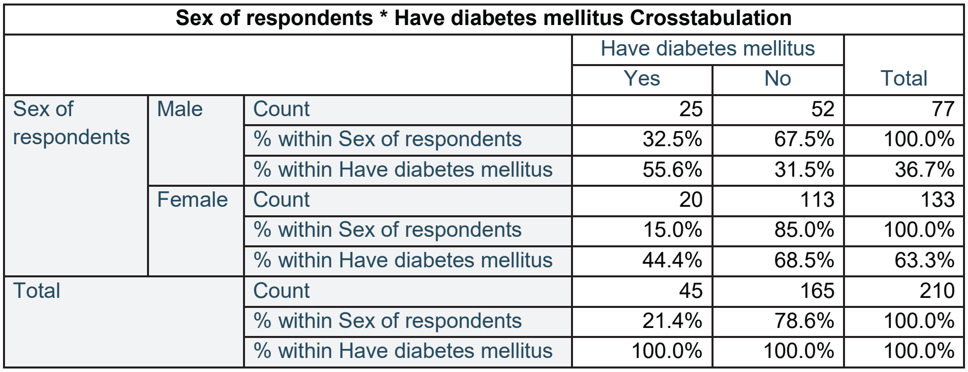
Percentages
- ☑ Row
- ☑ Column
- □ Total

Residuals
- □ Unstandardized
- □ Standardized
- □ Adjusted standardized

Noninteger Weights
- ● Round cell counts
- ○ Round case weights
- ○ Truncate cell counts
- ○ Truncate case weights
- ○ No adjustments

11.1.2 Outputs

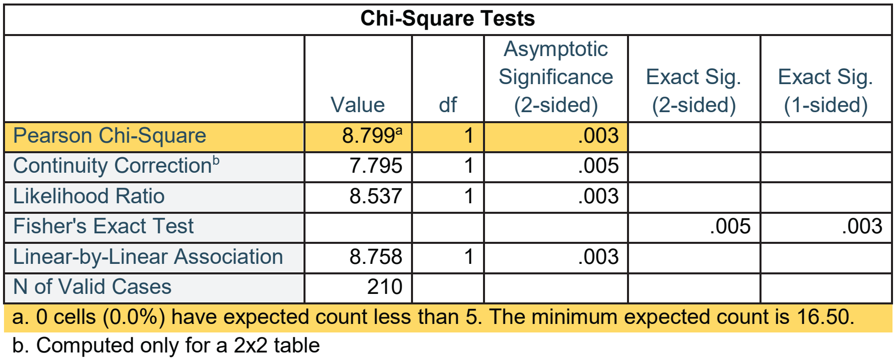
Table 11.1 Cross-tabulation of gender (sex) and diabetes mellitus



Sex of respondents * Have diabetes mellitus Crosstabulation

			Have diabetes mellitus		Total
			Yes	No	
Sex of respondents	Male	Count	25	52	77
		% within Sex of respondents	32.5%	67.5%	100.0%
		% within Have diabetes mellitus	55.6%	31.5%	36.7%
	Female	Count	20	113	133
		% within Sex of respondents	15.0%	85.0%	100.0%
		% within Have diabetes mellitus	44.4%	68.5%	63.3%
Total		Count	45	165	210
		% within Sex of respondents	21.4%	78.6%	100.0%
		% within Have diabetes mellitus	100.0%	100.0%	100.0%

Table 11.2 Chi-square test results with p-value



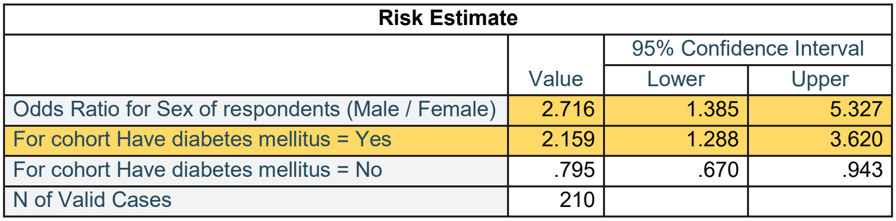
Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.799[a]	1	.003		
Continuity Correction[b]	7.795	1	.005		
Likelihood Ratio	8.537	1	.003		
Fisher's Exact Test				.005	.003
Linear-by-Linear Association	8.758	1	.003		
N of Valid Cases	210				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 16.50.

b. Computed only for a 2x2 table

Table 11.3 Odds Ratio (OR) and Relative Risk (RR) with 95% Confidence Interval (CI)



Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Sex of respondents (Male / Female)	2.716	1.385	5.327
For cohort Have diabetes mellitus = Yes	2.159	1.288	3.620
For cohort Have diabetes mellitus = No	.795	.670	.943
N of Valid Cases	210		

11.1.3 Interpretation

Table 11.1 is a 2 by 2 table of sex and diabetes with row (% within sex of respondents) and column (% within have diabetes mellitus) percentages. Now, the question is: which percentage should you report? It depends on the situation/study design and what do you want to report. For the data of a cross-sectional study, it may provide better information to the readers if row percentages are reported. In this situation, the row percentages will indicate the prevalence of the condition. For example, one can understand from Table 11.1 that the prevalence of diabetes among males is 32.5% and that of females is 15.0%, when row percentages are used. However, column percentages can also be reported for the cross-sectional data (most of the publications use column percentages). If column percentage is used, the meaning would be different. In this example (Table 11.1), it means that among those who have diabetes, 55.6% are male, compared to 31.5% who do not have diabetes. If the data are from a case-control study, you must report the column percentages (you cannot use row percentages for case-control studies). On the other hand, for the data of a cohort study, one should report the row percentages. In this case, it would indicate the incidence of the disease among males and females.

We can see in Table 11.1 (in the row of total) that the overall prevalence (irrespective of gender) of diabetes is 21.4% (consider that the data is from a cross-sectional study). Table 11.1 also shows that 32.5% of the males have diabetes compared to only 15.0% among females (i.e., the prevalence among males and females). Chi-square actually tests the hypothesis of whether the prevalence of diabetes among males and females is the same in the population or not.

Table 11.2 shows the Pearson Chi-square test results, including the degree of freedom (df) and p-value (Asymp. Sig). The table also shows other test results, such as Continuity Correction and *Fisher's Exact test*. Before we look at the Chi-square test results, it is important to check if there is any cell in the 2 by 2 table with an expected value of <5. This information is given at the bottom of Table 11.2 at "a" as "0 cells (0%) have expected count less than 5". For the use of the Chi-square test, it is desirable to have no cell in a 2 by 2 table with an expected count of less than 5. If this is not fulfilled, we have to use the Fisher's Exact test p-value to interpret the result (see Table 11.2). In fact, to use the Chi-square test, *no more than 20% cells* should have an expected frequency <5. You can get the expected frequencies for all the cells if you select "Expected" under "Count" in "Cell" option during analysis.

For the Chi-square test, consider the Pearson Chi-square value (Table 11.2) since there is no cell with an expected value <5. In our example, Chi-square value is 8.799 and the p-value is 0.003 (Table 11.2). Since the p-value is <0.05, there is an association between gender and diabetes. It can, therefore, be concluded that the prevalence of diabetes among males is significantly higher than that of females, which is statistically significant at 95%

confidence level (p=0.003).

Table 11.3 shows the OR (2.716) and its 95% CI (1.385-5.327). Use the OR if the data are from a case-control study. The OR is likewise used for cross-sectional data. The table also provided the RR (2.159) and its 95% CI (1.288-3.620) [consider the RR and 95% CI of the row "For cohort Have diabetes mellitus = Yes"]. Use RR if the data are from a cohort study. Note that both the OR and RR are statistically significant as they do not include 1 in the 95% CI. Odds ratio of 2.71 indicates that males are 2.7 times more likely to have diabetes compared to females. On the other hand, RR 2.15 indicates that the risk of having diabetes is 2.1 times higher in males compared to females. SPSS will *not* provide the OR and RR, if there are more than 2 categories in any of the variables (e.g., a 2 by 3 table). In such a situation, you have to use other ways to get the OR and RR (see section 11.1.4).

Table 11.4 Decision for using Chi-square test

Situation	Right test
Sample size >100 and expected cell value >10	Pearson's Chi-square (uncorrected)
Sample size 31-100 and expected cell count between 5-9	Pearson's Chi-square with Yate's correction (continuity correction row in Table 11.2)
Sample size less than 30 and/or expected cell value <5	Fisher's Exact test

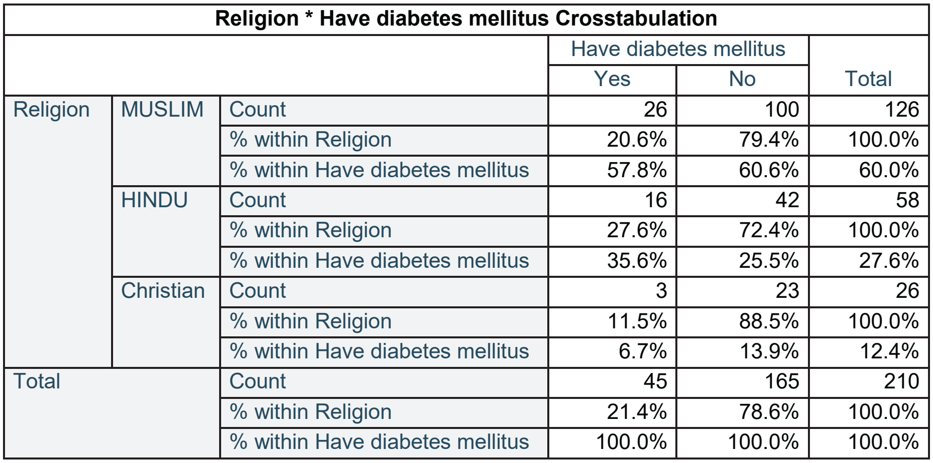
11.1.4 How to get OR and RR when the independent variable has more than two categories?

For example, we want to find association between diabetes (dependent variable) and religion (independent variable with 3 categories, 1= Muslim, 2= Hindu and 3= Christian) including the OR and RR. To find an association between these two variables, we shall use the following commands.

Analyze > Descriptive statistics > Crosstabs > Select "religion" and push it into the "Row(s)" box > Select "diabetes" for the "Column(s)" box > Statistics > Select "Chi-square" and "Risk" > Continue > Cells > Select "Raw" and "Column" under percentages > Continue > OK

With these commands, SPSS will produce the following tables (Tables 11.5 to 11.7).

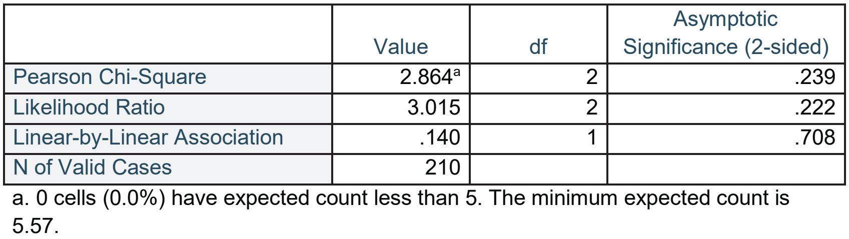
Table 11.5 Cross-tabulation of religion and diabetes



Religion * Have diabetes mellitus Crosstabulation

			Have diabetes mellitus		Total
			Yes	No	
Religion	MUSLIM	Count	26	100	126
		% within Religion	20.6%	79.4%	100.0%
		% within Have diabetes mellitus	57.8%	60.6%	60.0%
	HINDU	Count	16	42	58
		% within Religion	27.6%	72.4%	100.0%
		% within Have diabetes mellitus	35.6%	25.5%	27.6%
	Christian	Count	3	23	26
		% within Religion	11.5%	88.5%	100.0%
		% within Have diabetes mellitus	6.7%	13.9%	12.4%
Total		Count	45	165	210
		% within Religion	21.4%	78.6%	100.0%
		% within Have diabetes mellitus	100.0%	100.0%	100.0%

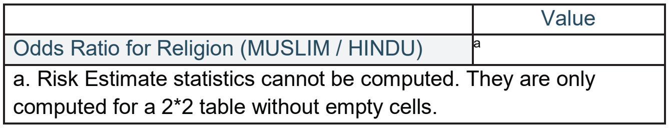
Table 11.6 Chi-Square Tests



	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2.864[a]	2	.239
Likelihood Ratio	3.015	2	.222
Linear-by-Linear Association	.140	1	.708
N of Valid Cases	210		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.57.

Table 11.7 Risk Estimate



	Value
Odds Ratio for Religion (MUSLIM / HINDU)	[a]

a. Risk Estimate statistics cannot be computed. They are only computed for a 2*2 table without empty cells.

Table 11.5 shows the frequency and percentage of different religious groups by diabetes. For example, the prevalence (% within religion) of diabetes among Muslims, Hindus and Christians is 20.6%, 27.6% and 11.5%, respectively (assume that the data are from a cross-sectional study). Though there are some differences in the prevalence of diabetes among the religious groups, there is no significant association between diabetes and religion since the p-value of the Pearson Chi-square test is 0.239 (>0.05) (Table 11.6).

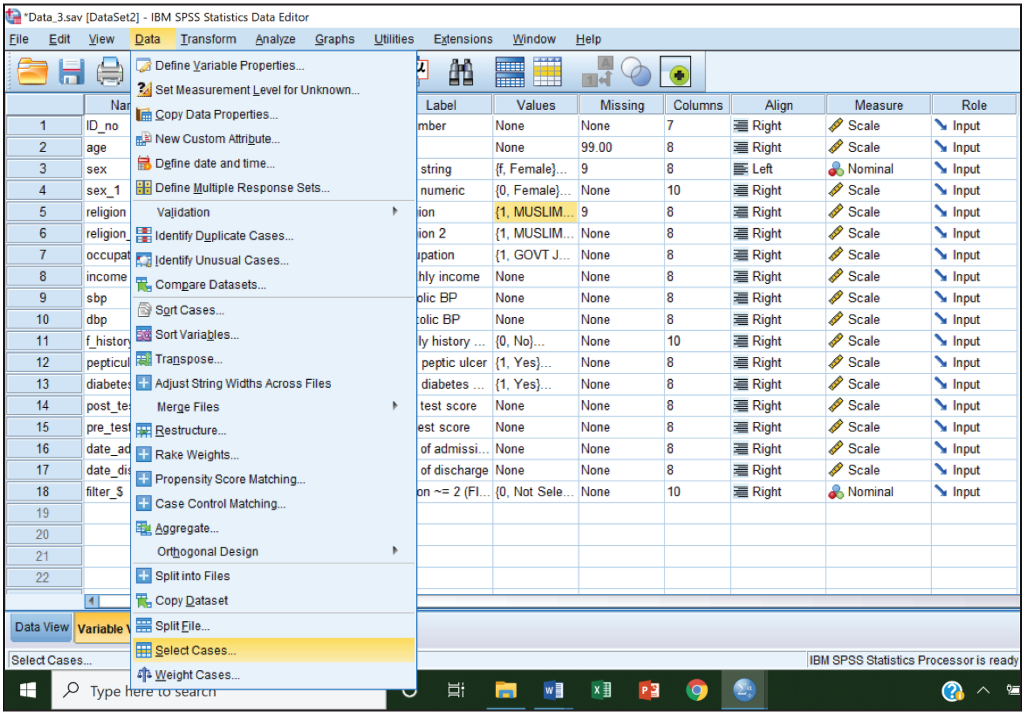
Table 11.7, which is supposed to provide the OR and RR, did not show any results. This is because there are more than 2 categories in the independent variable. To obtain the measures of risk (OR and RR) when there are 3 or more categories in the independent variable, first we shall have to decide about the comparison group. Let us select Christians as the comparison group. In that case, SPSS will provide the estimates of OR and RR for Muslims and Hindus compared to Christians.

Since we cannot get the risk estimates in a 2 by 3 table (2 categories of diabetes and 3 categories of religion), we have to first select 2 categories of religion (say, Muslim and Christian) for the analysis. This can be done using the following commands.

Data > Select cases > Select "If condition is satisfied" under "Select" > Click on "If" > Select "religion" and push it into the empty box > Click "~=" and then "2" from the "Number pad" > Continue > OK (Figs 11.5 to 11.7)

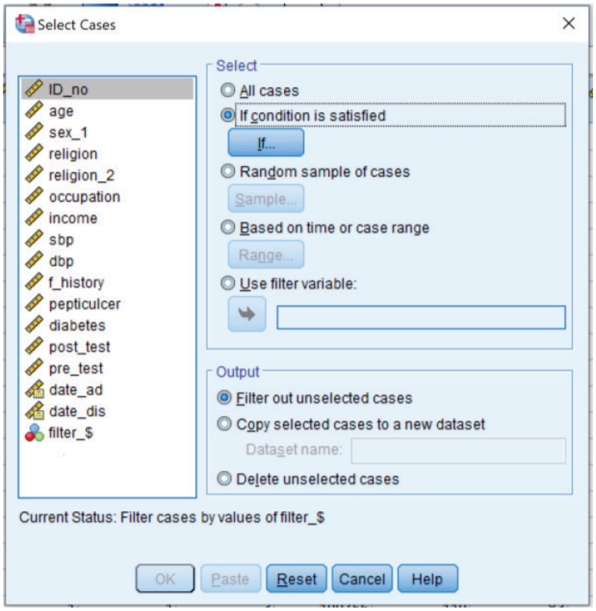
***Note**: The sign "~=" indicates "not equal to". Since we have selected 2 after this sign, SPSS will exclude code 2 (Hindus) from analysis.*

Figure 11.5



	Nar	Label	Values	Missing	Columns	Align	Measure	Role
1	ID_no	mber	None	None	7	Right	Scale	Input
2	age		None	99.00	8	Right	Scale	Input
3	sex	string	{f, Female}...	9	8	Left	Nominal	Input
4	sex_1	numeric	{0, Female}...	None	10	Right	Scale	Input
5	religion	ion	{1, MUSLIM...	9	8	Right	Scale	Input
6	religion_	ion 2	{1, MUSLIM...	None	8	Right	Scale	Input
7	occupat	ipation	{1, GOVT J...	None	8	Right	Scale	Input
8	income	hly income	None	None	8	Right	Scale	Input
9	sbp	olic BP	None	None	8	Right	Scale	Input
10	dbp	tolic BP	None	None	8	Right	Scale	Input
11	f_history	ly history ...	{0, No}...	None	10	Right	Scale	Input
12	pepticul	peptic ulcer	{1, Yes}...	None	8	Right	Scale	Input
13	diabetes	diabetes ...	{1, Yes}...	None	8	Right	Scale	Input
14	post_te	test score	None	None	8	Right	Scale	Input
15	pre_test	est score	None	None	8	Right	Scale	Input
16	date_ad	of admissi...	None	None	8	Right	Scale	Input
17	date_dis	of discharge	None	None	8	Right	Scale	Input
18	filter_$	on ~= 2 (FI...	{0, Not Sele...	None	10	Right	Nominal	Input
19								
20								
21								
22								

Figure 11.6



Select Cases

ID_no
age
sex_1
religion
religion_2
occupation
income
sbp
dbp
f_history
pepticulcer
diabetes
post_test
pre_test
date_ad
date_dis
filter_$

Select

○ All cases

● If condition is satisfied

If...

○ Random sample of cases

Sample...

○ Based on time or case range

Range...

○ Use filter variable:

Output

● Filter out unselected cases

○ Copy selected cases to a new dataset

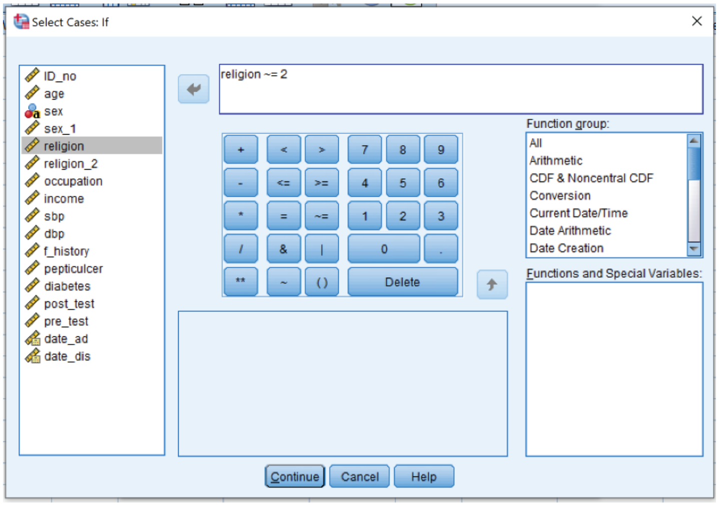
Dataset name:

○ Delete unselected cases

Current Status: Filter cases by values of filter_$

OK Paste Reset Cancel Help

Figure 11.7



Select Cases: If

ID_no
age
sex
sex_1
religion
religion_2
occupation
income
sbp
dbp
f_history
pepticulcer
diabetes
post_test
pre_test
date_ad
date_dis

religion ~= 2

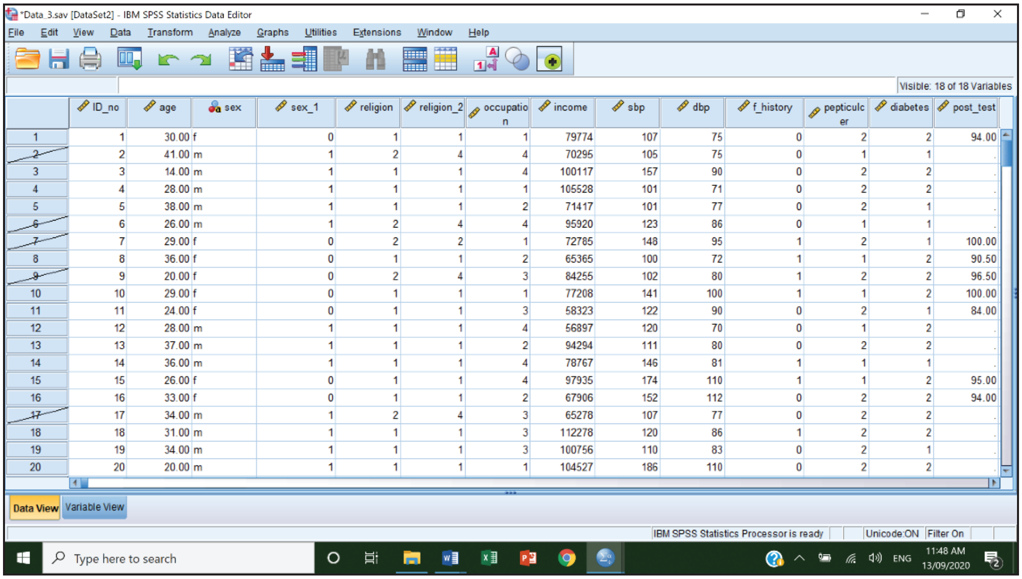
Function group:

All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation

Functions and Special Variables:

Continue Cancel Help

Figure 11.8


*Data_3.sav [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Visible: 18 of 18 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON
Filter On
Type here to search
ENG
11:48 AM
13/09/2020

	ID_no	age	sex	sex_1	religion	religion_2	occupation	income	sbp	dbp	f_history	pepticulcer	diabetes	post_test
1	1	30.00	f	0	1	1	1	79774	107	75	0	2	2	94.00
~~2~~	2	41.00	m	1	2	4	4	70295	105	75	0	1	1	.
3	3	14.00	m	1	1	1	4	100117	157	90	0	2	2	.
4	4	28.00	m	1	1	1	1	105528	101	71	0	2	2	.
5	5	38.00	m	1	1	1	2	71417	101	77	0	2	1	.
~~6~~	6	26.00	m	1	2	4	4	95920	123	86	0	1	1	.
~~7~~	7	29.00	f	0	2	2	1	72785	148	95	1	2	1	100.00
8	8	36.00	f	0	1	1	2	65365	100	72	1	1	2	90.50
~~9~~	9	20.00	f	0	2	4	3	84255	102	80	1	2	2	96.50
10	10	29.00	f	0	1	1	1	77208	141	100	1	1	2	100.00
11	11	24.00	f	0	1	1	3	58323	122	90	0	2	1	84.00
12	12	28.00	m	1	1	1	4	56897	120	70	0	1	2	.
13	13	37.00	m	1	1	1	2	94294	111	80	0	2	2	.
14	14	36.00	m	1	1	1	4	78767	146	81	1	1	1	.
15	15	26.00	f	0	1	1	4	97935	174	110	1	1	2	95.00
16	16	33.00	f	0	1	1	2	67906	152	112	0	2	2	94.00
~~17~~	17	34.00	m	1	2	4	3	65278	107	77	0	2	2	.
18	18	31.00	m	1	1	1	3	112278	120	86	1	2	2	.
19	19	34.00	m	1	1	1	3	100756	110	83	0	2	1	.
20	20	20.00	m	1	1	1	1	104527	186	110	0	2	2	.

Now, if you go to the data view mode of SPSS, you will see that some of the subjects have been excluded as indicated by "/" (back slash) (Fig 11.8). Besides this, you will not see any other changes. Now, to get the Chi-square test with risk estimates, use the following commands (same commands as used earlier).

> Analyze > Descriptive statistics > Crosstabs > Select "religion" and push it into the "Row(s)" box > Select "diabetes" for the "Column(s)" box > Statistics > Select "Chi-square" and "Risk" > Continue > Cells > Select "Raw" and "Column" under percentages > Continue > OK

SPSS will provide the following tables (Tables 11.8 to 11.10) along with the table for risk estimates (Table 11.10). You can notice that there is no information on Hindus in the cross-table (Table 11.8). Chi-square test results are provided in Table 11.9 (p-value of Pearson Chi-square is 0.282, which is not statistically significant). Table 11.10 shows the OR (OR= 1.99; 95% CI of OR: 0.55-7.15) and RR (RR= 1.78; 95% CI of RR: 0.58-5.47).

Table 11.8 Religion * Have diabetes mellitus Crosstabulation

			Have diabetes mellitus		Total
			Yes	No	
Religion	MUSLIM	Count	26	100	126
		% within Religion	20.6%	79.4%	100.0%
		% within Have diabetes mellitus	89.7%	81.3%	82.9%
	Christian	Count	3	23	26
		% within Religion	11.5%	88.5%	100.0%
		% within Have diabetes mellitus	10.3%	18.7%	17.1%
Total		Count	29	123	152
		% within Religion	19.1%	80.9%	100.0%
		% within Have diabetes mellitus	100.0%	100.0%	100.0%

Table 11.9 Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.155[a]	1	.282		
Continuity Correction[b]	.641	1	.423		
Likelihood Ratio	1.275	1	.259		
Fisher's Exact Test				.412	.216
Linear-by-Linear Association	1.148	1	.284		
N of Valid Cases	152				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.96.

b. Computed only for a 2x2 table

Table 11.10 Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Religion (MUSLIM / Christian)	1.993	.555	7.156
For cohort Have diabetes mellitus = Yes	1.788	.585	5.470
For cohort Have diabetes mellitus = No	.897	.761	1.058
N of Valid Cases	152		

Now, to get the risk estimate of Hindus compared to Christians, we shall have to exclude Muslims from the analysis using the following commands.

Data > Select cases > Select "If condition is satisfied" under "Select" > Click on "If" > You can see that in the box it is already "religion ~= 2" > Delete "2" and select "1" (1 is the code no. for Muslim) from the "Number pad" > Continue > OK

This will exclude Muslims from the analysis. Now, do the Chi-square test using the commands as shown above. You will get the OR and RR (including the 95% CI) for Hindus compared to Christians.

After completing this analysis, select “all cases” (Fig 11.6) for further analysis using the following commands.

Data > Select cases > Select “All cases” > OK (Fig 11.6)

12

Association between Two Continuous Variables: Correlation

Nature and strength of relationship between two or more continuous variables can be determined by *regression and correlation* analysis. *Correlation* is concerned with measuring the *strength of relationship* between continuous variables. The correlation model provides information on the relationship between two variables, without distinguishing which is dependent and which is independent variable. But the basic procedure for regression and correlation model is the same.

Under the correlation model, we calculate the "r" value. The "r" is called the sample *correlation coefficient*. It ("r" value) indicates the degree of linear relationship between dependent (Y) and independent (X) variable. The value of "r" ranges from +1 to –1. In this chapter, the correlation model is discussed. Use the data file <**Data_3.sav**> for practice.

12.1 Pearson's correlation

Pearson's correlation is used when the normality assumption is met [i.e., both the dependent and independent variables are normally distributed; assumption 1). For example, we want to explore if there is a correlation/association between systolic blood pressure (BP) (variable name is "sbp") and diastolic BP (variable name is "dbp").

Hypothesis

H_0: There is no correlation between systolic and diastolic BP.

H_A: There is a correlation between systolic and diastolic BP.

Assumption

1. The variables (systolic and diastolic BP) are normally distributed in the population;

2. The subjects represent a random sample from the population.

The first step, before doing the correlation analysis, is to generate a scatter diagram. The scatter diagram provides information/ideas about:

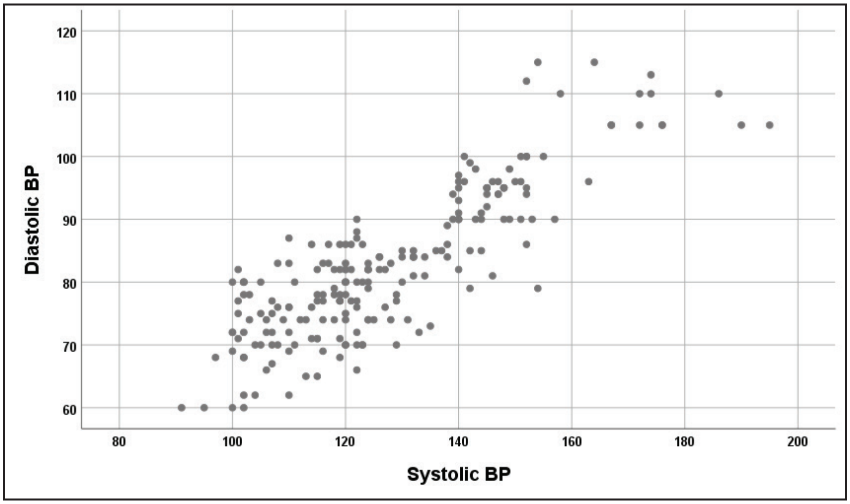
- Whether there is any correlation between the variables;
- Whether the relationship (if there is any) is linear or non-linear; and
- Direction of the relationship, i.e., whether it is positive (if the value of one varia--ble increases with the increase of the other variable) or negative (if the value of one variable decreases with the increase of the other variable).

12.1.1 Commands for scatter plot

To get the scatter plot of systolic and diastolic BP, use the following commands.

Graphs > Legacy dialogs > Scatter/Dot… > Select "Simple scatter" > Define > Select "sbp" for "X-axis" and "dbp" for "Y-axis" > Select "ID_no" for "Level cases by" (this would label the outliers, if there is any, by its ID number; you may omit it if you like) > OK

Figure 12.1 Scatter plot of systolic and diastolic BP



If you want to get the *regression line* on the scatter plot,

- Double click on the scatter plot in SPSS output
- You will see the template as shown in Figure 12.2 (Chart Editor)
- Click on "Elements" and then select "Fit Line at Total"
- You will see the template as shown in Figure 12.3

- Make sure that “Linear” is selected (if not, select “Linear”)
- Click on “Apply” and then on “Close” (or, on “Close” if “Apply” is not highlighted)
- Finally close the “Chart Editor” clicking at “×”

The SPSS will produce the scatter plot with the regression line on it (Fig 12.4). In the same manner, you can produce the scatter diagram of age and diastolic BP (Fig 12.5).

Figure 12.2 Chart Editor

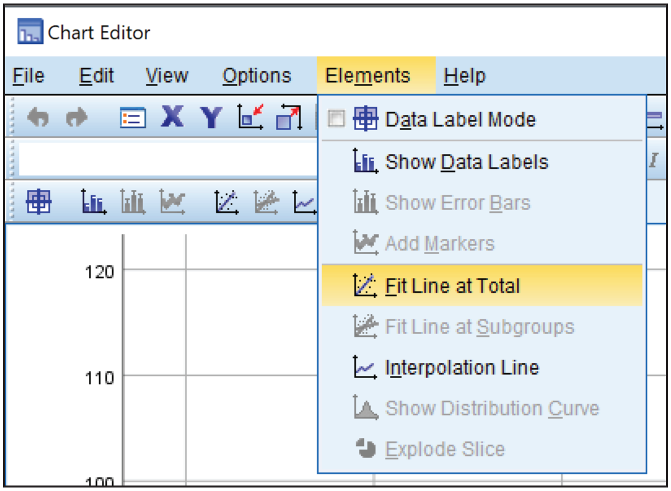
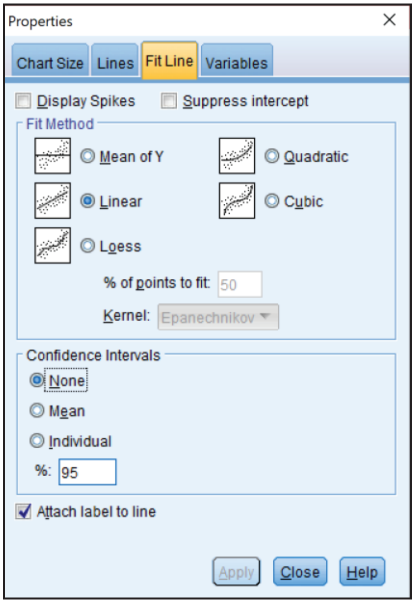


Figure 12.3 Chart Editor



□ Display Spikes □ Suppress intercept

Fit Method

○ Mean of Y ○ Quadratic

● Linear ○ Cubic

○ Loess

% of points to fit: 50

Kernel: Epanechnikov

Confidence Intervals

● None

○ Mean

○ Individual

%: 95

☑ Attach label to line

Figure 12.4 Scatter diagram of systolic and diastolic BP with regression line

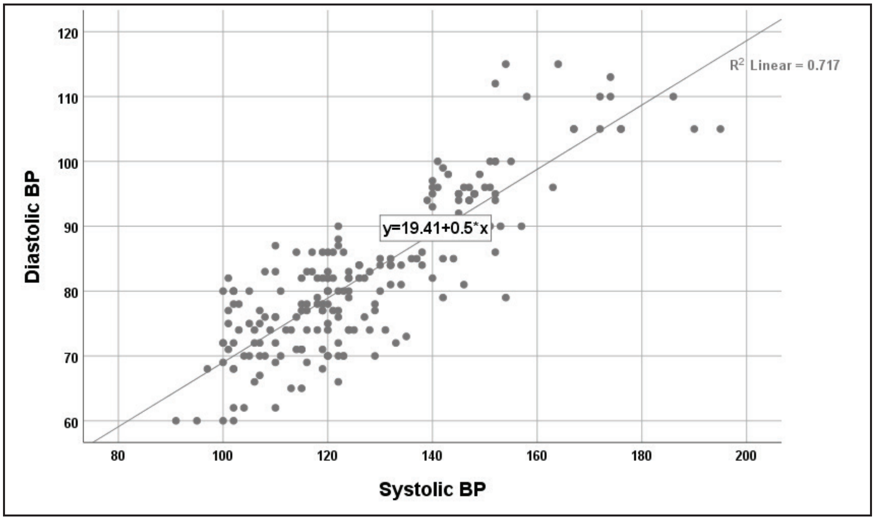
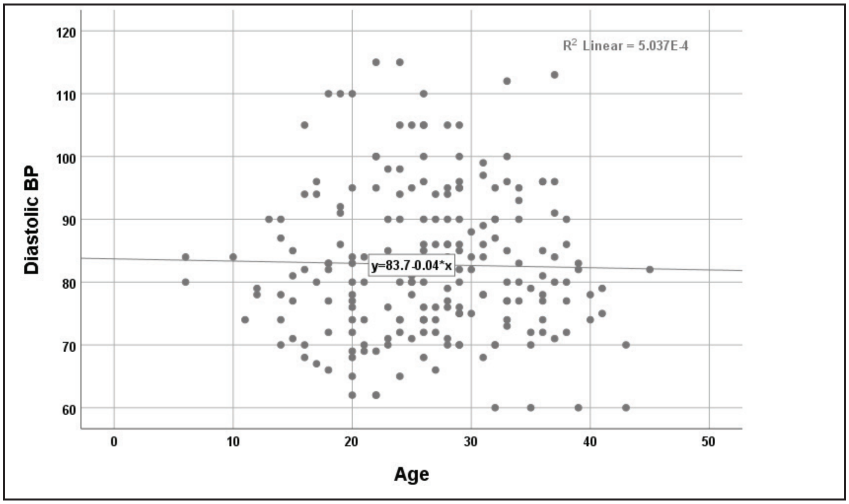


Figure 12.5 Scatter diagram of diastolic BP and age with regression line

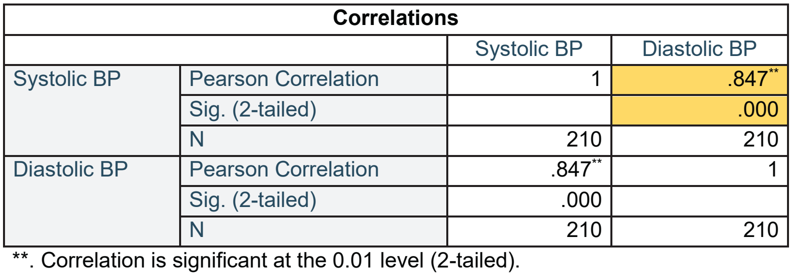


12.1.2 Commands for Pearson’s correlation

Analyze > Correlate > Bivariate > Select the variables “sbp” and “dbp” and push them into the "Variables" box > Select “Pearson” under the “Correlation coefficients” (usually set as default) > OK

12.1.3 Outputs

Table 12.1 Pearson's correlation between systolic and diastolic BP



Correlations

		Systolic BP	Diastolic BP
Systolic BP	Pearson Correlation	1	.847**
	Sig. (2-tailed)		.000
	N	210	210
Diastolic BP	Pearson Correlation	.847**	1
	Sig. (2-tailed)	.000	
	N	210	210

**. Correlation is significant at the 0.01 level (2-tailed).

12.1.4 Interpretation

In the first step, we have constructed the scatter plot of systolic and diastolic BP (Figs 12.1 and 12.2). Figure 12.1 shows that the data points are scattered around an invisible straight line and there is an increase in the diastolic BP (Y) as the systolic BP increases (X). This indicates that there may have a positive correlation between these two variables. Look at Figure 12.2, which shows the regression line in the scatter plot. The regression line has passed from near to the lower left corner to the upper right corner, indicating a positive correlation between systolic and diastolic BP. If the relationship were negative (inverse), the regression line would have passed from the upper left corner to the lower right corner. Figure 12.3 shows the scatter plot of diastolic BP and age. It does not indicate any correlation between diastolic BP and age, since the dots are scattered around the regression line, which is more or less parallel to the X-axis.

For correlation, look at the value of Pearson's correlation. Table 12.1 shows that the correlation coefficient (Pearson Correlation) of systolic and diastolic BP is 0.847 and the p-value is 0.000. Correlation coefficient [r] indicates the *strength/degree of linear relationship* between the two variables (systolic and diastolic BP). As the value of "r" is positive and the p-value is <0.05, there is a significant positive correlation between systolic and diastolic BP.

The value of "r" lies between –1 and +1. Values near to "zero" indicate no correlation, while values near to "+1" or "–1" indicate a strong correlation. The negative value of "r" (– r) indicates an inverse relationship. A value of r ≥ 0.8 indicates a very strong correlation; "r" value between 0.6 and 0.8 indicates a moderately strong correlation; "r" value between 0.3 and 0.5 indicates a fair correlation and "r" value <0.3 indicates a poor correlation.

12.2 Spearman's correlation

Spearman's correlation (instead of Pearson's correlation) is used when the normality ass-

-umption is violated (i.e., if the distribution of either the dependent or independent or both the variables are non-normal). Spearman's correlation is also applicable for two categorical ordinal variables, such as intensity of pain (mild, moderate and severe pain) and grade of cancer (stage 1, stage 2 and stage 3).

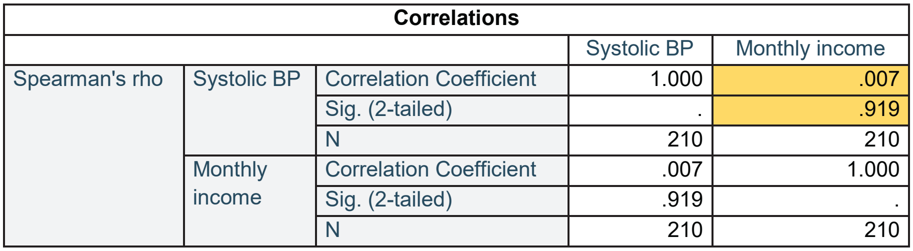
Suppose we want to determine if there is a correlation between systolic BP (variable name is "sbp") and income, where income is not normally distributed in the population. Spearman's correlation is the appropriate statistical method to test the hypothesis in this scenario.

12.2.1 Commands for Spearman's correlation

Analyze > Correlate > Bivariate > Select the variables "sbp" and "income" and push them into the "Variables" box > Select "Spearman" under the "Correlation coefficients" > OK

12.2.2 Outputs

Table 12.2 Spearman's correlation between systolic BP and income



Correlations

			Systolic BP	Monthly income
Spearman's rho	Systolic BP	Correlation Coefficient	1.000	.007
		Sig. (2-tailed)	.	.919
		N	210	210
	Monthly income	Correlation Coefficient	.007	1.000
		Sig. (2-tailed)	.919	.
		N	210	210

12.2.3 Interpretation

Table 12.2 shows the Spearman's correlation between systolic BP and income. The results indicate that there is no correlation between systolic BP and income (r= 0.007; p=0.919), since the "r" value (correlation coefficient) is very small and the p-value is >0.05.

12.3 Partial correlation

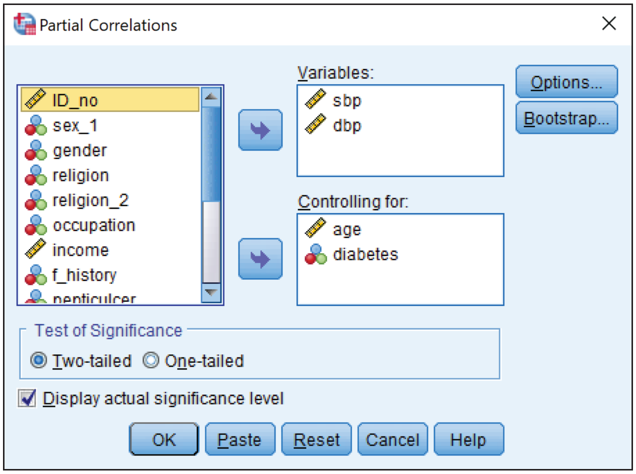
The purpose of doing the partial correlation is to assess the correlation (indicated by the "r" value) between two variables after adjusting/controlling for one or more other variables (continuous or categorical). This means that through partial correlation, we obtain the adjusted "r" value after controlling for the confounding factors. For example, if we assume that the relationship between systolic and diastolic BP may be influenced (confounded) by other variables (such as age and diabetes), we should do the partial correlati-

-on to exclude the influence of other variables (age and diabetes). The partial correlation will provide the correlation coefficient ("r" value) between systolic and diastolic BP after controlling/ adjusting for age and diabetes.

12.3.1 Commands for partial correlation

> Analyze > Correlate > Partial > Select "sbp" and "dbp" for "Variables" box > Select "age" and "diabetes" for "Controlling for" box > OK (Fig 12.6)

Figure 12.6



Variables:
sbp
dbp

Controlling for:
age
diabetes

Test of Significance
● Two-tailed ○ One-tailed

☑ Display actual significance level

12.3.2 Outputs

Table 12.3 Correlation between systolic and diastolic BP after controlling for age and diabetes mellitus

Correlations				
Control Variables			Systolic BP	Diastolic BP
Age & Have diabetes mellitus	Systolic BP	Correlation	1.000	.847
		Significance (2-tailed)	.	.000
		df	0	206
	Diastolic BP	Correlation	.847	1.000
		Significance (2-tailed)	.000	.
		df	206	0

12.3.3 Interpretation

Table 12.3 shows the results of partial correlation between systolic and diastolic BP after adjusting/controlling for age and diabetes mellitus. We can see in the table that r=0.847 and p=0.000. This means that these two variables (systolic and diastolic BP) are signific-

-antly correlated (p=0.000), even after controlling for age and diabetes mellitus.

If the relationship between systolic and diastolic BP were influenced by age and diabetes mellitus, the crude (unadjusted) and adjusted “r” values would be different. Look at Table 12.1 (Pearson’s correlation), which shows the crude “r” value (r=0.847). After adjusting for age and diabetes (Table 12.3), the “r” value remains the same (r=0.847). Since the crude and adjusted “r” values are same (or close to each other), there is no influence of age and diabetes mellitus in the relationship between systolic and diastolic BP (i.e., age and diabetes mellitus are not the confounding factors in the relationship between systolic and diastolic BP).

13

Linear Regression

Regression analysis is a commonly used statistical method for data analysis. Nature and strength of relationship between two or more continuous variables can be determined by regression and correlation analysis. We have already discussed correlation in the previous chapter. While *correlation* is concerned about measuring the *direction and strength of linear relationship* between the variables, *regression* analysis is helpful to *predict or estimate* the value of one variable corresponding to a value of another variable(s) (e.g., to understand whether systolic BP is a good predictor of diastolic BP). In regression analysis, our main interest is in *regression coefficient* (also called slope or β), which indicates the strength of association between dependent (Y) and independent (X) variables. Regression can be done as: *a) Simple linear regression, and b) Multiple linear regression methods.*

In this chapter, both simple and multiple linear regressions are discussed. Multiple linear regression is a type of *multivariable analysis*. Multivariable analysis is a statistical tool where multiple independent variables are considered for a single outcome (variable). The terms "multivariate analysis" and "multivariable analysis" are often used interchangeably in health research. Multivariate analysis actually refers to the statistical method for the analysis of multiple outcomes.

Multivariable analyses are widely used in observational studies, intervention studies (randomized and non-randomized trials), and studies of diagnosis and prognosis. The main purposes of multivariable analysis are to:

a) Determine the relative contribution of independent variables to the outcome variable;
b) Adjust for the confounding factors;
c) Predict the probability of an outcome, when several characteristics are present in an individual; and
d) Assess interaction of multiple variables for the outcome.

There are several types of multivariable analysis methods. The choice of multivariable analysis for the type of outcome variable, is summarized in Table 7.3 (chapter 7). The co-

-mmonly used multivariable analysis methods in health research include multiple linear regression, logistic regression and proportional hazards regression (Cox regression) that are discussed in this book. Use the data file <Data_4.sav> for practice.

13.1 Simple linear regression

In simple linear regression, there is one dependent and one independent variable. The objective of simple linear regression is to find the *population regression equation*, which describes the true relationship between the dependent variable (Y) and independent variable (X). In simple linear regression model, two variables are involved – one is independent variable (X), placed on X-axis, and the other is dependent variable (Y), placed on Y-axis. Then, we call it "regression of Y on X".

Suppose we want to perform a simple linear regression analysis of diastolic BP (dependent variable) on systolic BP (independent variable). The objective is to find the population regression equation to predict the diastolic BP by systolic BP.

Assumptions

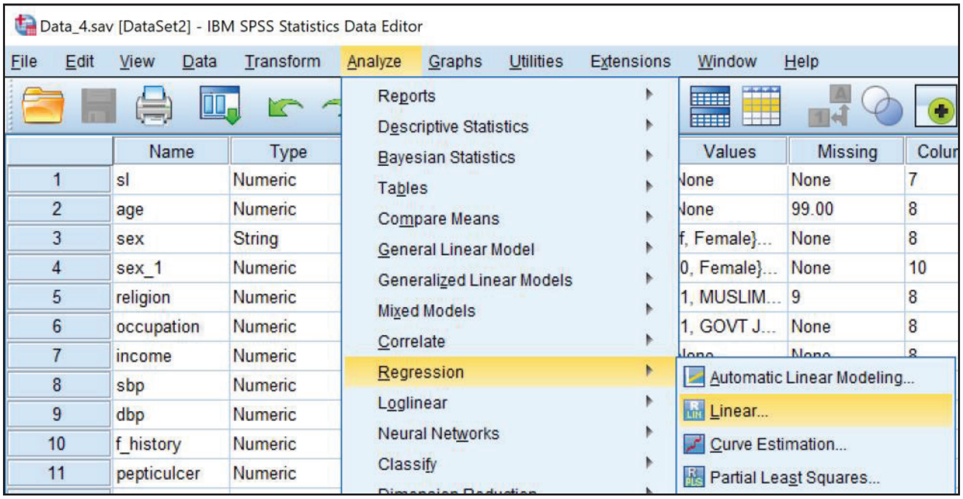
1. **Normality:** For any fixed value of X (systolic BP), the sub-population of Y values (diastolic BP) is normally distributed;
2. **Homoscedasticity:** The variances of the sub-populations of "Y" are all equal;
3. **Linearity:** The means of the sub-populations of "Y" lie on the same straight line;
4. **Independence:** Observations are independent of each other.

The first step in analysing the data for regression is to construct a scatter diagram, as discussed in chapter 12. This will provide an indication of the linear relationship between the variables, systolic and diastolic BP.

13.1.1 Commands

Analyze > Regression > Linear > Select "dbp" for "Dependent" box and "sbp" for "Independent(s)" box > Method "Enter" (usually the default) > Statistics > Select "Estimates, Descriptive, Confidence interval, and Model fit" > Continue > Options > Select "Exclude cases pairwise" under "Missing values" > Continue > OK (Figs 13.1 to 13.3)

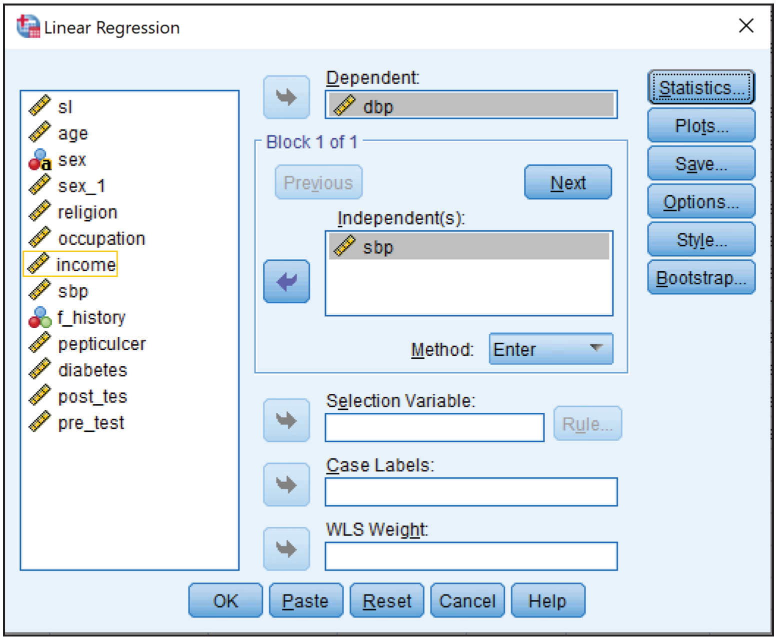
Figure 13.1



Data_4.sav [DataSet2] - IBM SPSS Statistics Data Editor

	Name	Type	Values	Missing	Colur
1	sl	Numeric	None	None	7
2	age	Numeric	None	99.00	8
3	sex	String	f, Female}...	None	8
4	sex_1	Numeric	0, Female}...	None	10
5	religion	Numeric	1, MUSLIM...	9	8
6	occupation	Numeric	1, GOVT J...	None	8
7	income	Numeric	None	None	8
8	sbp	Numeric			
9	dbp	Numeric			
10	f_history	Numeric			
11	pepticulcer	Numeric			

Figure 13.2



Linear Regression

Dependent: dbp

Independent(s): sbp

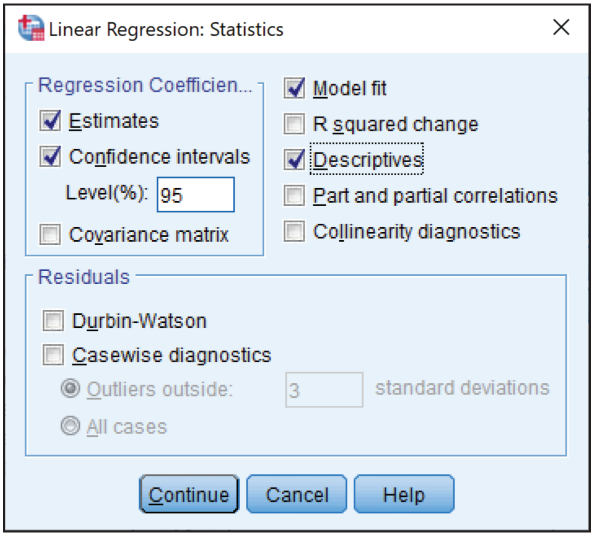
Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Figure 13.3



Linear Regression: Statistics

Regression Coefficien...

☑ Estimates

☑ Confidence intervals

Level(%): 95

☐ Covariance matrix

☑ Model fit

☐ R squared change

☑ Descriptives

☐ Part and partial correlations

☐ Collinearity diagnostics

Residuals

☐ Durbin-Watson

☐ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

13.1.2 Outputs

Table 13.1 Mean and standard deviation of the variables

Descriptive Statistics			
	Mean	Std. Deviation	N
Diastolic BP	82.77	11.749	210
Systolic BP	127.73	20.058	210

Table 13.2 Correlation between systolic and diastolic BP

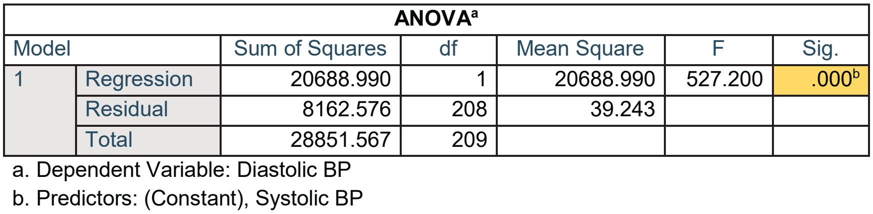
Correlations			
		Diastolic BP	Systolic BP
Pearson Correlation	Diastolic BP	1.000	.847
	Systolic BP	.847	1.000
Sig. (1-tailed)	Diastolic BP	.	.000
	Systolic BP	.000	.
N	Diastolic BP	210	210
	Systolic BP	210	210

Table 13.3 Correlation coefficient (R) and coefficient of determination (R-square)

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.847[a]	.717	.716	6.264

a. Predictors: (Constant), Systolic BP

Table 13.4 ANOVA table for significance of “R”

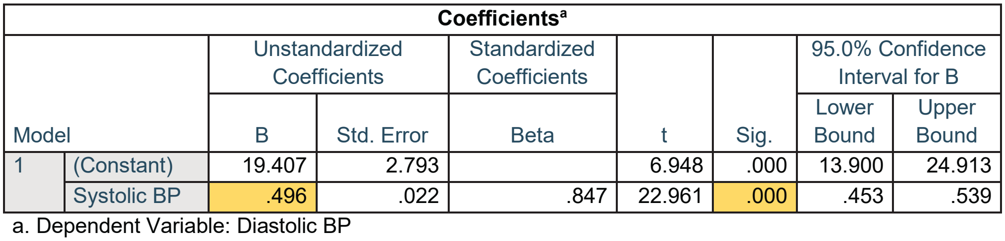


ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20688.990	1	20688.990	527.200	.000[b]
	Residual	8162.576	208	39.243		
	Total	28851.567	209			

a. Dependent Variable: Diastolic BP

b. Predictors: (Constant), Systolic BP

Table 13.5 Constant (a) and regression coefficient (b)



Coefficients[a]								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	19.407	2.793		6.948	.000	13.900	24.913
	Systolic BP	.496	.022	.847	22.961	.000	.453	.539

a. Dependent Variable: Diastolic BP

13.1.3 Interpretation

Tables 13.1 and 13.2 provide the descriptive statistics (mean and standard deviation) and correlation coefficient (“r” value, which is 0.847) of the diastolic and systolic BP. The model summary table (Table 13.3) shows the Pearson’s correlation coefficient “R” (r = 0.847) and coefficient of determination “R-square” ($r^2 = 0.717$).

It is important to note the value of R-square (coefficient of determination) given in the model summary table (Table 13.3). R-square indicates the amount of variation in “Y” due to “X”, that can be explained by the regression line. Here, the R-square value is 0.717 (~0.72), which indicates that 72% variation in diastolic BP can be explained by the systolic BP. The rest of the variation (28%) is due to other factors (unexplained variation). The adjusted R-square value (0.716), as shown in the table, is the value when the R-square is adjusted for better population estimation.

The ANOVA table (Table 13.4) indicates whether the correlation coefficient (R) is significant or not (i.e., whether the linear regression model is useful to explain the dependent variable by the independent variable). As the p-value (Sig.) of the F-test (F=527.2) is 0.000, R is statistically significant at 95% confidence level. We can, therefore, conclude that there is a significant positive (because R value is positive) correlation between the diastolic and systolic BP, and we can use the regression equation for prediction. The table also shows the regression (also called explained) sum of squares (23890.586) and residual (also called error) sum of squares (8528.028). The residual indicates the difference between the observed value and predicted value (i.e., the observed value and the value on the regression line). The residual sum of squares provides an idea about how well the regression line actually fits into the data. The smaller the value better is the fit.

Table 13.5 (coefficients) provides quantification of the relationship between the diastolic and systolic BP. The table shows the values for "a" (constant or Y-intercept) and "b" (unstandardized coefficients - B) or slope (also called regression coefficient, β). Note that for a single independent variable, the standardized coefficient (Beta) is equal to the Pearson's correlation value.

In our example, the value of "a" (constant) is 19.407 and "b" (unstandardized coefficient - B) is 0.496 (both are positive). The value, a= +19.407, indicates that the regression line crosses/cuts the Y-axis above the origin (zero) and at the point 19.407 (a negative value indicates that the regression line cuts the Y-axis below the origin). This value (value for "a") does not have any practical meaning, since it indicates the average diastolic BP of individuals, if the systolic BP is 0 (zero).

The value of "b (unstandardized coefficient)" (the regression coefficient or slope) indicates the amount of variation/change in "Y" (here it is diastolic BP) for each unit change in "X" (systolic BP). Here, the value of "b" is 0.496, which means that if the systolic BP increases (or decreases) by 1 mmHg, the diastolic BP will increase (or decrease) by 0.496 mmHg. The table also shows the significance (p-value) of "b", which is 0.000. A p-value <0.05 indicates that "b" is not equal to zero in the population. Note that for simple linear regression, if R is significant, "b" will also be significant and will have the same sign (positive or negative).

We know that the simple linear regression equation is, $Y = a + bX$ ("Y" is the predicted value of the dependent variable; "a" is the Y-intercept or constant; "b" is the regression coefficient and "X" is a value of the independent variable). Therefore, the regression/prediction equation for this regression model is:

$$Y = 19.407 + 0.496 \times X.$$

With this equation, we can estimate the diastolic BP by the systolic BP. For example, what would be the estimated diastolic BP of an individual whose systolic BP is 130 mmHg? The answer is, the estimated diastolic BP would be equal to (19.407 + 0.496 × 130) 83.89 mmHg.

Note that, if we want to use the regression equation for the purpose of prediction/estimation, "b" has to be statistically significant ($p<0.05$). In our example, the p-value for "b" is 0.000, and we can, therefore, use the equation for the prediction of diastolic BP by systolic BP.

The analysis (Table 13.5) has actually evaluated whether "b" in the population is "zero" or not by the t-test (*Null hypothesis*: the regression coefficient (b) is equal to "zero" in the population; *Alternative hypothesis*: the population regression coefficient is not equal to "zero"). We can reject the null hypothesis, since the p-value is <0.05. It can, therefore, be concluded that the systolic BP can be used to predict/estimate the diastolic BP using the regression equation, Y = 19.407 + 0.496*X.

13.2 Multiple linear regression

In simple linear regression, two variables are involved - one dependent (Y) and one independent (X) variable. The independent variable is also called *explanatory or predictor* variable. In multiple regression, there are more than one explanatory (independent) variables in the model. The explanatory variables may be quantitative or categorical. The main purposes of multiple regression analysis are to:

- Obtain the adjusted estimates of regression coefficients (B) of the explanatory variables in the model;
- Predict or estimate the value of the dependent variable by the explanatory variables in the model; and
- Understand the amount of variation in the dependent variable explained by the explanatory variables in the model.

For instance, we want to assess the contribution of four variables (age, systolic BP, sex and religion) in explaining the diastolic BP in a sample of individuals selected randomly from a population. Here, the dependent variable is the diastolic BP and the explanatory variables (independent variables) are age, systolic BP, sex and religion. Of the explanatory (independent) variables, two are quantitative (age and systolic BP) and two are categorical variables (sex and religion). Of the categorical variables, sex has two levels (male and female) and religion has three levels (Muslim, Hindu and Christian). When the independent variable is categorical with more than two levels (e.g., religion), we need to create dummy variables for that variable. For example, if we want to include the variable "religion" in the regression model, we need to create dummy variables for religion.

13.2.1 Creating dummy variables

In our example, the variable "religion" has three levels and are coded as 1= Muslim; 2= Hindu and 3= Christian. We cannot simply put "religion" as one of the explanatory variables in the regression model, because the coding is arbitrary and the regression estimates obtained for religion would be meaningless. We need to generate dummy variables for religion.

The number of dummy variables to be generated for "religion" is two (no. of levels minus 1). Before generating the dummy variables, we need to decide about the comparison group. Let us consider "Christian" as the comparison group, and assign "0" (zero) as its code number. We shall generate two dummy variables – one is "reli_1" and the other is "reli_2" for religion. To generate the dummy variables, we shall have to recode the variable "religion" using the following commands.

Transform > Create dummy variables > Select "religion" and push it into "Create dummy variables for" box > Write "reli" in the "Root names (One per selected

The above commands will generate three dummy variables – "reli_1" (coded as 1= Muslim; 0= other religions, i.e., Hindu and Christian); "reli_2" (coded as 1= Hindu; 0= other religions, i.e., Muslim and Christian); and "reli_3" (coded as 1= Christian; 0= other religions, i.e., Muslim and Hindu) and are labelled as "religion=Muslim", "religion=Hindu" and "religion=Christian", respectively. You will find all the dummy variables at the bottom of the variable view of the data file. Provide the value labels of all the dummy variables as discussed in section 2.1.1.

Since we have decided "Christian" as the comparison group and we need two dummy variables, we shall use the dummy variables "reli_1 (religion=Muslim)" and "reli_2 (religion=Hindu)" during the analysis.

You can also generate the dummy variables one-by-one separately using the following commands.

Step 1: Create the first dummy variable "**reli_1**" for religion

Transform > Recode into different variables > Select "religion" and push it into the "Input variable – Output variable" box > Write "reli_1" in the "Name" box under "Output variable" > In the "Label" box write "Muslim" > Change > Click on "Old and New Values.." > Select "Value" under "Old value" and write 1 in the box > Select "Value" under "New value" and write 1 in the box > Add > Select "All other values" under "Old value" > Write 0 (zero) in the box "Value" under the "New value" > Add > Continue > OK

Step 2: Create the second dummy variable "**reli_2**" for religion

Transform > Recode into different variables > Select "religion" and push it into the "Input variable – Output variable" box > Write "reli_2" in the "Name" box under "Output variable" > In the "Label" box write "Hindu" > Change > Click on "Old and New Values.." > Select "Value" under "Old value" and write 2 in the box > Select "Value" under "New value" and write 1 in the box > Add > Select "All other values" under "Old value" > Write 0 (zero) in the box "Value" under the "New value" > Add > Continue > OK

The above commands will create two dummy variables for religion, the "reli_1" (for which code 1= Muslim and 0= other religions, i.e., Hindu and Christian)" and "reli_2" (for which code 1= Hindu and 0= other religions, i.e., Muslim and Christian)". You can see the new variables at the bottom of the variable view of the data file. *Don't forget to provide the value labels for the dummy variables.*

13.2.2 Changing string variables to numeric variables

If we want to include the variable “sex” in the model, we need to check its coding. If the variable is coded as string variable (e.g., m= male and f= female, as is done in our data), we need to recode it as a numeric variable, say, 0= female and 1= male. In this case, when multiple regression will be performed, the regression estimate in the model will be for males compared to females (the lower value will be the comparison group). Use the following commands to generate a numeric variable for sex (you can also use the other option as discussed in section 6.2).

> Transform > Recode into different variables > Select “sex” and push it into the “Input variable – Output variable” box > Write “sex_1” in the “Name” box under “Output variable” > Write “Sex numeric” in the “Label” box > Click on “Change” > Click on “Old and New Values..” > Select “Value” under “Old value” and write **f** in the box > Select “Value” under “New value” and type **0** in the box > Add > Select “Value” under “Old value” and write **m** in the box > Write **1** in the box “Value” under the “New value” > Add > Continue > OK

This will generate a new variable “sex_1” (last variable in the variable view) with codes 0= female and 1= male. Go to the “variable view” of the data file and set these code numbers in the column “Value” of the variable sex_1.

13.2.3 Sample size for multiple regression

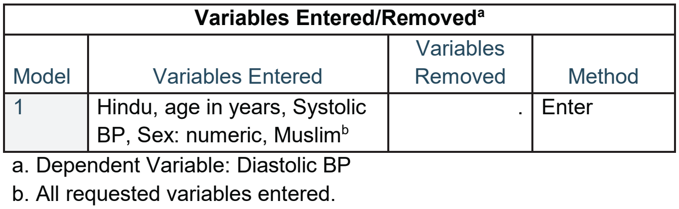
Multiple regression should be done if the sample size is fairly large. The minimum sample size needed for the analysis depends on how many independent variables we want to include in the model. Different authors provided different guidelines regarding this. One author recommended a minimum of 20 subjects for each of the independent variables in the model[14]. Another author provided a formula ($n \geq 50 + 8m$) to estimate the number of subjects required for the model[24]. Here “m” indicates the number of predictors. For example, if we intend to include 5 independent variables in the model, we need to have at least 90 subjects (50 + 8*5). For stepwise regression, there should be 40 cases for each of the independent variables in the model.

13.2.4 Commands for multiple linear regression

Use the following commands for multiple regression analysis, where the dependent variable is dbp (diastolic BP) and the explanatory (independent) variables are age, sbp (systolic BP), sex_1, reli_1 and reli_2.

Analyze > Regression > Linear > Select “dbp” for “Dependent" box and “age, sbp, sex_1, reli_1 and reli_2” for "Independent(s)” box > Method “Enter” (usually the default) > Statistics > Select “Estimates, Confidence interval, and Model fit” > Continue > Options > Select “Exclude cases pairwise” under “Missing values” > Continue > OK

Table 13.6 Variables entered (and removed) into the model



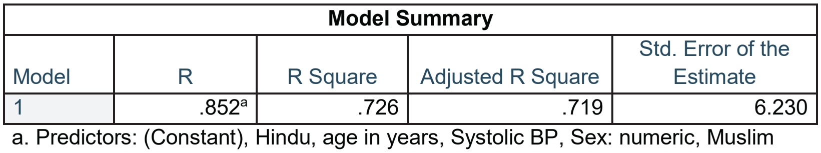
Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Hindu, age in years, Systolic BP, Sex: numeric, Muslim[b]	.	Enter

a. Dependent Variable: Diastolic BP

b. All requested variables entered.

Table 13.7 Multiple R, R-square and adjusted R-square values

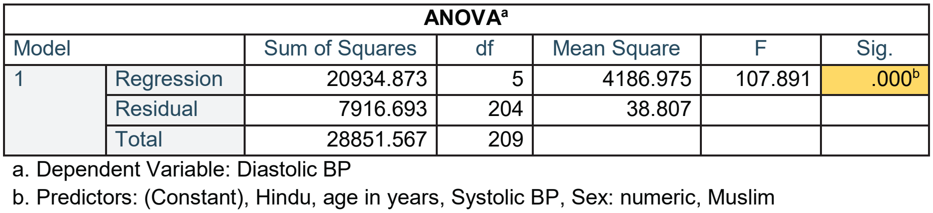


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.852[a]	.726	.719	6.230

a. Predictors: (Constant), Hindu, age in years, Systolic BP, Sex: numeric, Muslim

Table 13.8 ANOVA table for significance of R



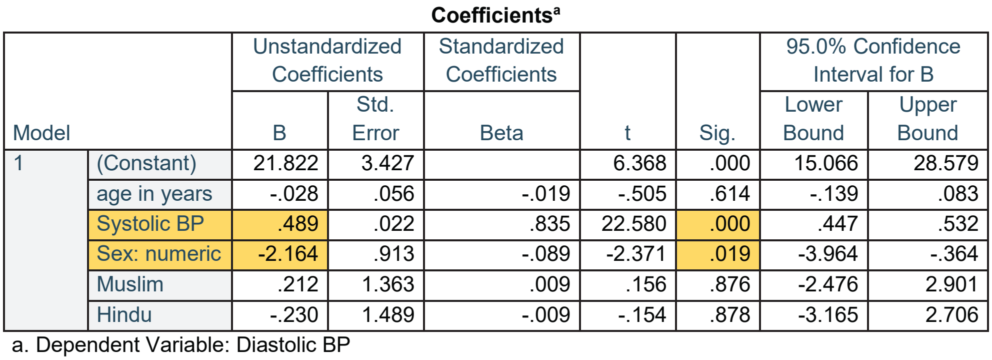
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20934.873	5	4186.975	107.891	.000[b]
	Residual	7916.693	204	38.807		
	Total	28851.567	209			

a. Dependent Variable: Diastolic BP

b. Predictors: (Constant), Hindu, age in years, Systolic BP, Sex: numeric, Muslim

Table 13.9 Adjusted regression coefficients of explanatory variables and their significance



Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	21.822	3.427		6.368	.000	15.066	28.579
	age in years	-.028	.056	-.019	-.505	.614	-.139	.083
	Systolic BP	.489	.022	.835	22.580	.000	.447	.532
	Sex: numeric	-2.164	.913	-.089	-2.371	.019	-3.964	-.364
	Muslim	.212	1.363	.009	.156	.876	-2.476	2.901
	Hindu	-.230	1.489	-.009	-.154	.878	-3.165	2.706

a. Dependent Variable: Diastolic BP

13.2.6 Interpretation

Table 13.6 shows the variables that have been entered into and removed from the model. Since we have used the "Enter" method, SPSS has not removed any variable during analysis.

Table 13.7 (model summary) shows the values for R (0.852), R-square (0.726) and adjusted R-square (0.719) [adjusted for better population estimation]. In multiple regression, the R measures the correlation between the observed values of the dependent variable and the predicted values based on the regression model. The R-square may overestimate the population value, if the sample size is small. The adjusted R-square gives the R-square value for better population estimation. The R-square value of 0.726 indicates that all the independent variables (age, systolic BP, sex and religion) together in the model explains 72.6% variation in diastolic BP, which is statistically significant (p=0.000), as shown in the ANOVA table (Table 13.8).

The coefficients table (Table 13.9) shows regression coefficients (unstandardized and standardized), p-values (Sig.) and 95% confidence intervals (CI) for regression coefficients of all the explanatory variables in the model along with the constant. This is the most important table for interpretation of results. The unstandardized regression coefficients (B) are for age (0.028; p=0.614), systolic BP (0.489; p<0.001), sex (-2.164; p=0.019 for males compared to females), Muslims (0.212; p=0.876 compared to Christians) and Hindus (-0.230; p=0.878 compared to Christians).

From this output (Table 13.9), we conclude that the systolic BP and sex are the factors significantly influencing the diastolic BP (since the p-values are <0.05). The other variables in the model (age and religion) do not have any significant influence in explaining the diastolic BP. The unstandardized coefficient (B) [also called *multiple regression coefficient*] for systolic BP, in this example, is 0.489 (95% CI: 0.44 to 0.53). This indicates that the average increase (or decrease) in diastolic BP is 0.49 mmHg, if the systolic BP increases (or decreases) by 1 mmHg after adjusting for all other variables (age, sex and religion) in the model. On the other hand, the unstandardized coefficient (B) for sex is -2.164 (95% CI: -3.964 to -0.364), which indicates that (on an average) the diastolic BP of males is 2.2 mmHg less (since the coefficient is negative) than the females after adjusting for all other variables in the model. If the standardized coefficient were positive (i.e., +2.164), the diastolic BP (on an average) of males would have been 2.2 mmHg higher than the females, given the other variables constant in the model.

The standardized coefficients (Beta) (Table 13.9) indicate which independent variables have more influence on the dependent variable (diastolic BP). The bigger the value more is the influence. We can see in Table 13.9 that the standardized coefficients for systolic BP and sex are 0.835 and -0.089, respectively. This means that systolic BP has greater influence in explaining the diastolic BP than sex.

13.2.7 Regression equation

The regression equation to estimate the average value of the dependent variable with the explanatory variables is:

$$Y= a + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 \ldots\ldots.. B_nX_n$$

Here, "Y" is the estimated mean value of the dependent variable; "a" is the constant (or Y-intercept); "B" is the regression coefficient(s) of the variables in the model and "X" is the value of the variable(s) in the model.

Suppose we want to estimate the average diastolic BP of an individual who is 40 years old, male, Muslim and has systolic BP 120 mmHg. In Table 13.9 (coefficients), we can find the regression coefficients (unstandardized coefficients, B) for age [=0.028 (B_1)], systolic BP [=0.489 (B_2)], sex [= -2.164 (B_3) for being male] and Muslim [=0.212 (B_4) for being Muslim] and the constant (=21.822). Therefore, the estimated diastolic BP of the individual will be:

$$Y= 21.822 + (0.028 \times 40) + (0.489 \times 120) + (-2.164 \times 1) + (0.212*1) = \mathbf{79.67}.$$

13.2.8 Problem of multicollinearity

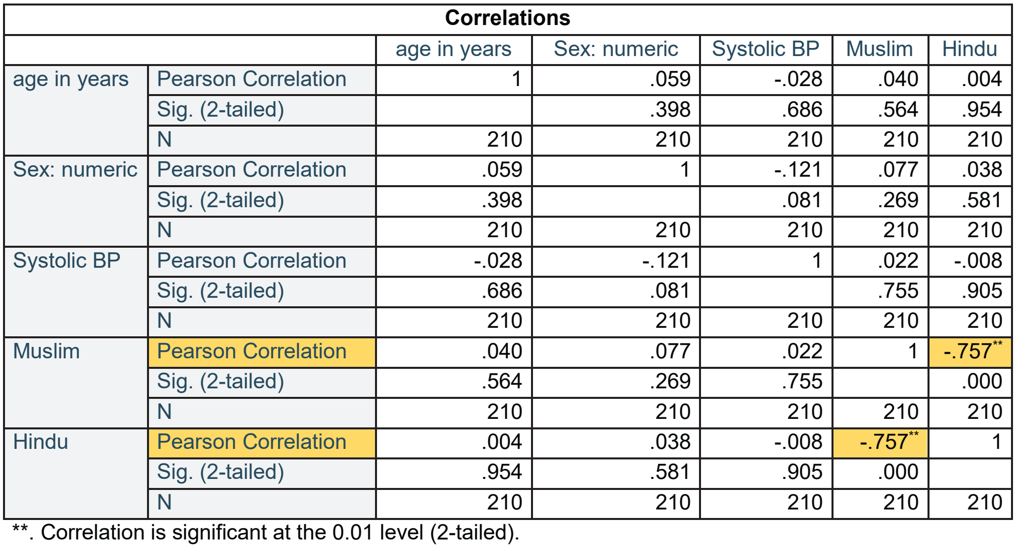
Before deciding about the multiple regression model, we need to check for *multicollinearity* (inter-correlations among the independent variables) of the independent variables. If there are moderate to high inter-correlations among the independent variables, two situations may occur. *Firstly*, the importance of a given explanatory variable may be difficult to determine because of biased (distorted) p-value; and the *other* is dubious relationships may be obtained. For example, if there is multicollinearity, we may observe that the regression coefficient for sex is not significant and the systolic BP has a negative relationship with the diastolic BP. Another important sign of multicollinearity is a *severe reduction of the Adjusted R Square value*.

To determine the correlations among the independent variables, we can generate the Pearson's correlation matrix. For example, we want to see the correlations among the systolic BP, age, sex and religion. Use the following commands to get the correlation matrix.

Analyze > Correlate > Bivariate... > Select the variables "age, sex_1, sbp, reli_1 and reli_2" for the "Variables" box > OK

The SPSS will produce the following correlation matrix table (Table 13.10).

Table 13.10 Correlation matrix of independent variables



Correlations

		age in years	Sex: numeric	Systolic BP	Muslim	Hindu
age in years	Pearson Correlation	1	.059	-.028	.040	.004
	Sig. (2-tailed)		.398	.686	.564	.954
	N	210	210	210	210	210
Sex: numeric	Pearson Correlation	.059	1	-.121	.077	.038
	Sig. (2-tailed)	.398		.081	.269	.581
	N	210	210	210	210	210
Systolic BP	Pearson Correlation	-.028	-.121	1	.022	-.008
	Sig. (2-tailed)	.686	.081		.755	.905
	N	210	210	210	210	210
Muslim	Pearson Correlation	.040	.077	.022	1	-.757**
	Sig. (2-tailed)	.564	.269	.755		.000
	N	210	210	210	210	210
Hindu	Pearson Correlation	.004	.038	-.008	-.757**	1
	Sig. (2-tailed)	.954	.581	.905	.000	
	N	210	210	210	210	210

**. Correlation is significant at the 0.01 level (2-tailed).

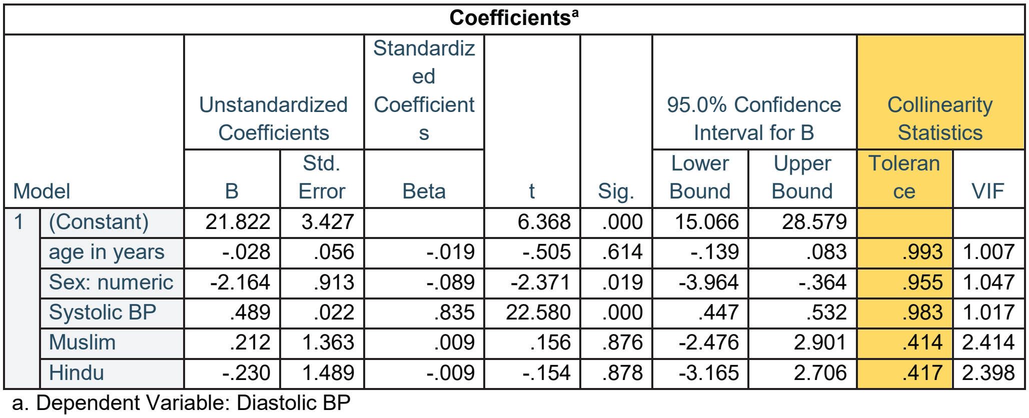
Table 13.10 shows that there is a moderately strong correlation (r = -0.757) between Muslim (reli_1) and Hindu (reli_2), while the correlation coefficients for other variables are low. However, the correlation between reli_1 and reli_2 (the dummy variables) did not affect our regression analysis.

Pearson's correlation can only check collinearity between any two variables. Sometimes a variable may be multicollinear with a combination of variables. In such a situation, it is better to use the tolerance measure (another measure for multicollinearity), which gives the strength of the linear relationships among the independent variables (usually the dummy variables have higher correlation). To get the *tolerance measure*, use the following commands.

Analyze > Regression > Linear > Select "dbp" for "Dependent" box and "age, sbp, sex_1, reli_1 and reli_2" for "Independent(s)" box > Method "Enter" > Statistics > Select "Estimates, Confidence interval, Model fit, and **Collinearity diagnostics**" > Continue > Options > Select "Exclude cases pairwise" under "Missing values" > Continue > OK

This will provide the collinearity statistics in the coefficients table as shown in Table 13.11.

Table 13.11 Collinearity statistics for multicollinearity diagnostics



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	21.822	3.427		6.368	.000	15.066	28.579		
	age in years	-.028	.056	-.019	-.505	.614	-.139	.083	.993	1.007
	Sex: numeric	-2.164	.913	-.089	-2.371	.019	-3.964	-.364	.955	1.047
	Systolic BP	.489	.022	.835	22.580	.000	.447	.532	.983	1.017
	Muslim	.212	1.363	.009	.156	.876	-2.476	2.901	.414	2.414
	Hindu	-.230	1.489	-.009	-.154	.878	-3.165	2.706	.417	2.398

a. Dependent Variable: Diastolic BP

The tolerance value ranges from 0 to 1. A value close to “zero” indicates that the variable is almost in a linear combination (i.e., has strong correlation) with other independent variables. In our example (Table 13.11), the tolerance values for age, sex, and systolic BP are more than 0.95. However, the tolerance values for Muslim (reli_1) and Hindu (reli_2) [the dummy variables] are a little more than 0.40. *The recommended tolerance level is more than 0.6 before we put the variable in the multiple regression model.* However, a tolerance of 0.40 and above is acceptable, especially if it is a dummy variable. The other statistic provided in the last column of the table is the VIF (Variance Inflation Factor). This is the inverse of the tolerance value.

If there are variables that are highly correlated (tolerance value is <0.4), one way to solve the problem is to exclude one of the correlated variables from the model. The other way is to combine the explanatory variables together (e.g., taking their sum).

Finally, for developing a model for multiple regression, we should first check for multicollinearity and then the residual assumptions (see below). If they fulfil the requirements, then only we can finalize the regression model.

13.2.9 Checking for assumptions

For practical purposes, there are three assumptions that need to be checked on the residuals for the linear regression model to be valid. The assumptions are:

a. There is no outlier;
b. The data points are independent;
c. The residuals are normally distributed with mean = 0 and have constant variance.

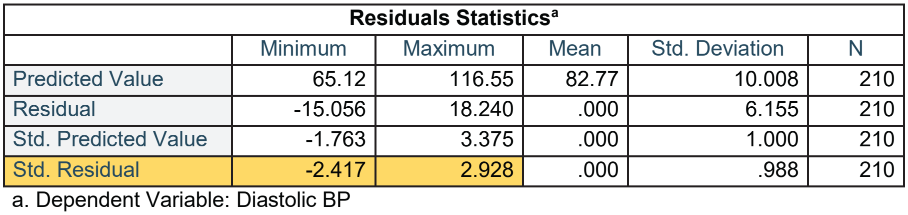
13.2.9.1 Checking for outliers and independent data points (assumptions a and b)

Use the following commands to check for outliers (casewise diagnostics) and data points are independent (Durbin-Watson statistics).

Analyze > Regression > Linear > Select “dbp” for “Dependent" box and “age, sex_1, sbp, reli_1 and reli_2” for "Independent(s)” box > Method “Enter” > Statistics > Select “Estimates, Confidence interval, Model fit, **Casewise diagnostics and Durbin-Watson**” > Continue > Options > Select “Exclude cases pairwise” under “Missing values” > Continue > OK

SPSS will produce the Residuals Statistics table (Table 13.12), Model Summary table (Table 13.13) and Casewise Diagnostics table (Table 13.14), if there are outliers, otherwise not.

Table 13.12 Residuals statistics without outliers in the data set

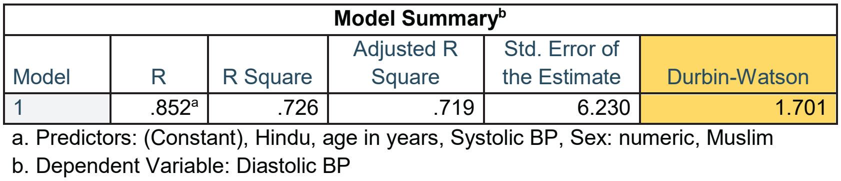


Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	65.12	116.55	82.77	10.008	210
Residual	-15.056	18.240	.000	6.155	210
Std. Predicted Value	-1.763	3.375	.000	1.000	210
Std. Residual	-2.417	2.928	.000	.988	210

a. Dependent Variable: Diastolic BP

Table 13.13 Durbin-Watson statistics for checking data points are independent



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.852[a]	.726	.719	6.230	1.701

a. Predictors: (Constant), Hindu, age in years, Systolic BP, Sex: numeric, Muslim

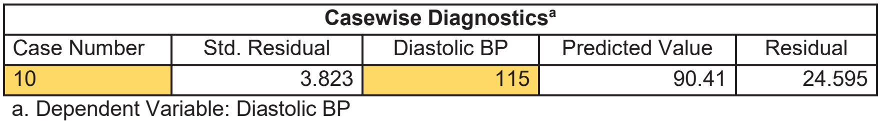
b. Dependent Variable: Diastolic BP

Look at the residuals statistics table (Table 13.12). Our interest is in the *Std. Residual* value. *The “minimum” and “maximum” values should not exceed “+3” or “–3”*. Table 13.12 shows that the minimum value is “–2.417” and the maximum value is “+2.928” (none of them exceeded “–3” or “+3”). This means that there are no outliers in the dependent variable. Since there are no outliers, SPSS did not provide the casewise diagnostics table.

The Durbin-Watson test is done to check whether data points are independent. The Model Summary table (Table 13.13) shows the Durbin-Watson statistic in the last column. *The Durbin-Watson estimate ranges from 0 to 4. Values around 2 indicate that the data points are independent.* Values near zero indicate a strong positive correlation and values near 4 indicate a strong negative correlation. If the data points are independent, the Durbin-Watson value hovers around 2. The table shows that the value of the Durbin-Watson statistic is 1.701, which is close to 2. We, therefore, conclude the data points are independent.

We have provided two additional tables below – Table 13.14 (Casewise Diagnostics) and Table 13.15 (Residual Statistics) from another dataset where outliers are present. Look at Table 13.14 (Casewise Diagnostics). The table shows that there is an outlier in the diastolic BP, the value of which is 115 and the case (ID) number is 10 (if there is no outlier in the data, this table will not be provided by SPSS). Now, look at the Residuals Statistics table (Table 13.15). The table shows that the standard residual value (3.823) has exceeded +3. This indicates that there are outliers in the dependent variable.

Table 13.14 Case number of the outlier

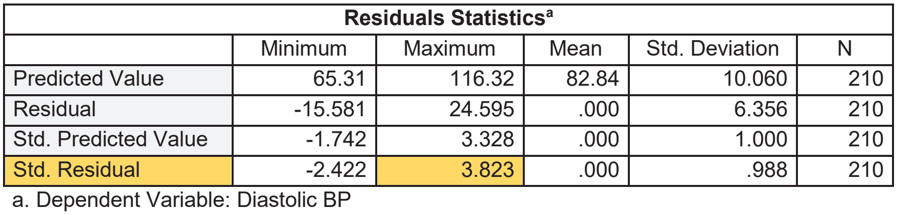


Casewise Diagnostics[a]

Case Number	Std. Residual	Diastolic BP	Predicted Value	Residual
10	3.823	115	90.41	24.595

a. Dependent Variable: Diastolic BP

Table 13.15 Residuals statistics with outliers in the dataset



Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	65.31	116.32	82.84	10.060	210
Residual	-15.581	24.595	.000	6.356	210
Std. Predicted Value	-1.742	3.328	.000	1.000	210
Std. Residual	-2.422	3.823	.000	.988	210

a. Dependent Variable: Diastolic BP

13.2.9.2 Checking for normality assumption of the residuals and constant variance

To check the normality of the residuals and constant variance, use the following commands.

Analyze > Regression > Linear > Select "dbp" for "Dependent" box and "age, sex_1, sbp, reli_1 and reli_2" for "Independent(s)" box > Method "Enter" > Statistics > Select "Estimates, Confidence interval, Model fit" > Continue > Options > Select "Exclude cases pairwise" under "Missing values" > Continue > Plots > Select "**Histogram and normal probability plot**" under "Standardized residual plots" > Place "***ZRESID in Y; and *ZPRED in X**" box under "Scatter 1 of 1" > Continue > OK

The SPSS will produce the histogram (Fig 13.4), normal probability (P-P) plot (Fig 13.5) and a scatter plot (Fig 13.6) of the residuals. The distribution of the residuals is normal as seen in the histogram (Fig 13.4) and P-P plot (Fig 13.5). The constant variance (homoscedasticity) is checked in the scatter plot (Fig 13.6). If the scatter of the points shows no clear pattern (as seen in Fig 13.6), we can conclude that the variances of the sub-population of "Y" are constant.

Figure 13.4 Histogram

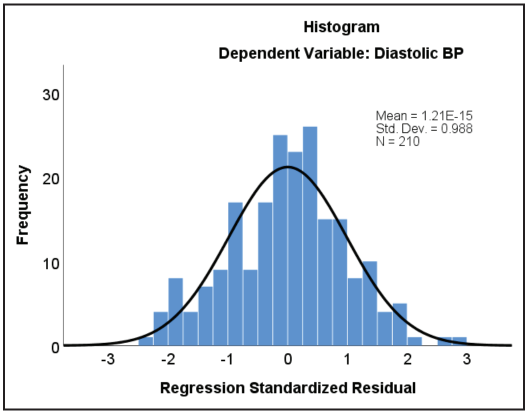


Figure 13.5 Normal probability plot

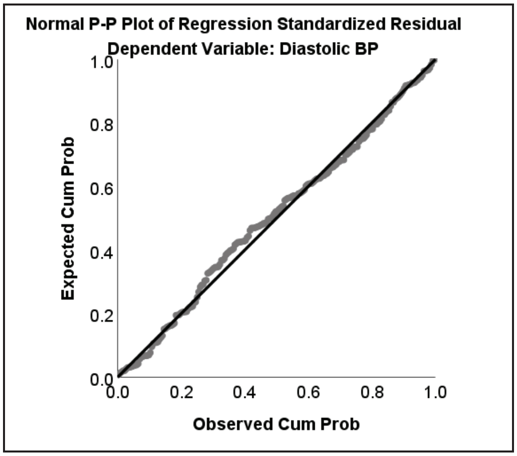
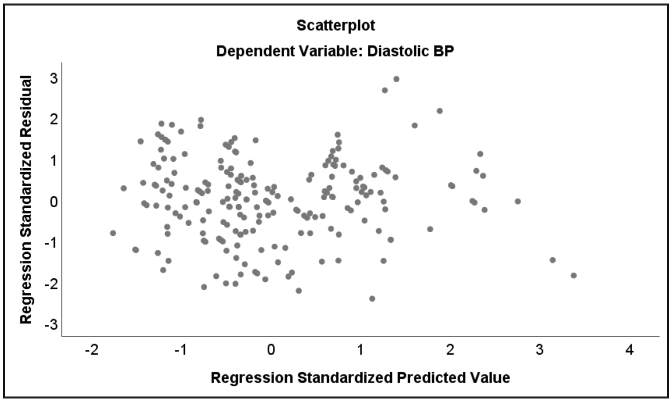


Figure 13.6 Scatter plot of standardized residual vs standardized predicted value



13.2.10 Variable selection for the model

In general, independent variables to be selected for multivariable analysis should include the risk factors of interest and potential confounders (based on theory, prior research and empirical findings), while variables with lots of missing values should be excluded.

We have used the "Enter" method for the analysis of data (modelling) earlier in this chapter. The "Enter" method uses all the independent variables in the model included by the researcher. It does not exclude any variable automatically from the model during analysis. Automatic procedures can be used to determine which independent variables will be included in the model. The main reason for using the automatic selection procedure is to minimize the number of independent variables necessary to estimate or predict the outcome. SPSS and other data analysis software have the option to automatically select the independent variables for the model. They use statistical criteria to select the variables and their order in the model. The commonly used variable selection techniques are provided in Table 13.16.

Let us see how to use the "Stepwise" method (commonly used method in multiple regression analysis) for modelling. To do this, use the following commands (only change is in "Method" selection).

Analyze > Regression > Linear > Select "dbp" for "Dependent" box and "age, sbp, sex_1, reli_1 and reli_2" for "Independent(s)" box > Method "**Stepwise**" (Fig 13.7) > Statistics > Select "Estimates, Confidence interval, and Model fit" > Continue > Options > Select "Exclude cases pairwise" under "Missing values" > Continue > OK

Note: If you click on "Option", you can see the criteria for entry and removal of variables (Fig 13.8).

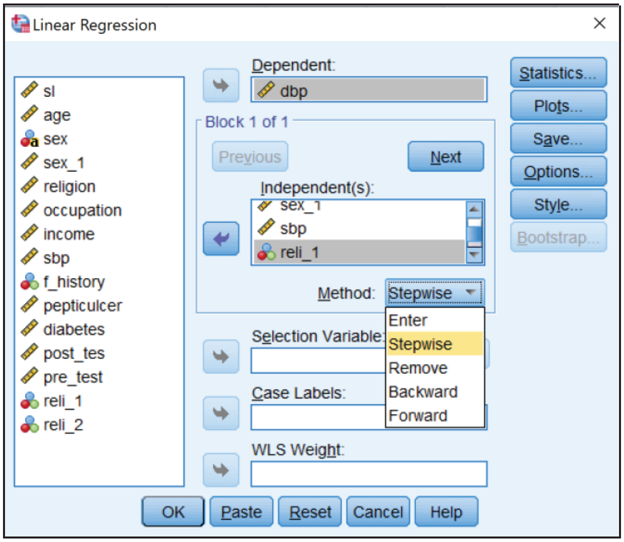
Table 13.16 Methods of variable selection

Techniques	Methods	Advantages and limitations
Forward	This method *enters* variables in the model sequentially. The order is determined by the variable's association (significance) with the outcome variable (variables with strongest association are entered first) after adjustment for the other variables already in the model.	Best suited for dealing with the studies where the sample size is small. Does not deal well with suppressor (confounding) effects.

Table 13.16 Methods of variable selection

Techniques	Methods	Advantages and limitations
Backward	This technique *removes* variables from the model sequentially. The order is determined by the variable's association with the outcome variable (variables with weakest association leave first) after adjustment for the variables already in the model.	Better for assessing suppressor effect than the forward selection method.
Stepwise/ Remove	This is the *combination* of forward and backward methods. In the stepwise method, variables that are entered are checked at each step for removal. Likewise, in the removal method, variables that are excluded will be checked for re-entry.	Has the ability to manage large number of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options.
Enter (all variables)	This method enters all the variables at the same time (does not remove any variable automatically from the model).	Including all variables may be problematic, if there are many independent variables and the sample size is small.

Figure 13.7 Model selection options

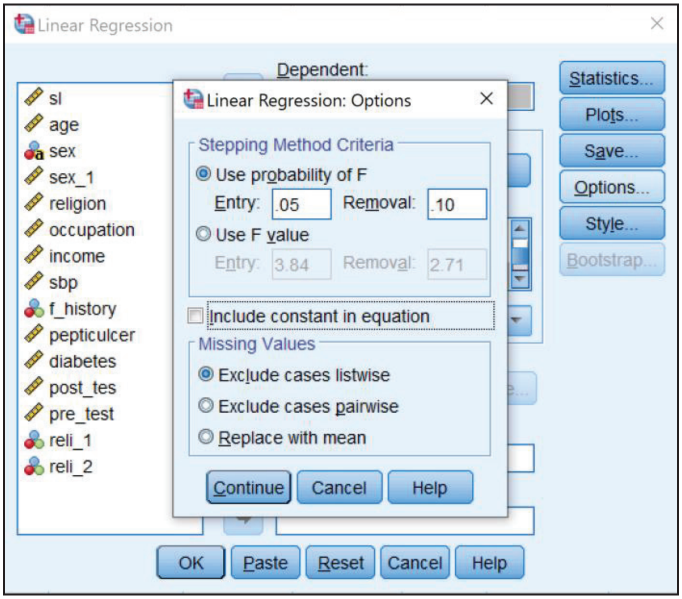


Dependent: dbp

Independent(s): sex_1, sbp, reli_1

Method: Stepwise

Figure 13.8 Selection criteria for entry and removal of variables



Linear Regression: Options

Stepping Method Criteria

● Use probability of F

Entry: .05 Removal: .10

○ Use F value

Entry: 3.84 Removal: 2.71

□ Include constant in equation

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

13.2.10.1 Outputs (only the relevant table is provided)

Table 13.17 Models and adjusted regression coefficients of independent variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	19.407	2.793		6.948	.000	13.900	24.913
	Systolic BP	.496	.022	.847	22.961	.000	.453	.539
2	(Constant)	21.016	2.838		7.405	.000	15.421	26.611
	Systolic BP	.490	.022	.836	22.768	.000	.447	.532
	Sex: numeric	-2.180	.893	-.090	-2.441	.015	-3.941	-.419

a. Dependent Variable: Diastolic BP

Table 13.18 Multiple R, R-square and adjusted R-square values by model

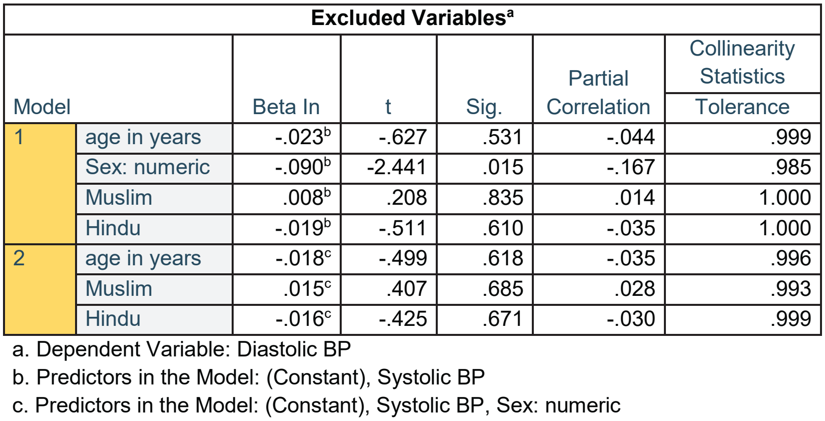
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.847[a]	.717	.716	6.264
2	.851[b]	.725	.722	6.191

a. Predictors: (Constant), Systolic BP

b. Predictors: (Constant), Systolic BP, Sex: numeric

Table 13.19 Variables removed from the models



Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	age in years	-.023[b]	-.627	.531	-.044	.999
	Sex: numeric	-.090[b]	-2.441	.015	-.167	.985
	Muslim	.008[b]	.208	.835	.014	1.000
	Hindu	-.019[b]	-.511	.610	-.035	1.000
2	age in years	-.018[c]	-.499	.618	-.035	.996
	Muslim	.015[c]	.407	.685	.028	.993
	Hindu	-.016[c]	-.425	.671	-.030	.999

a. Dependent Variable: Diastolic BP

b. Predictors in the Model: (Constant), Systolic BP

c. Predictors in the Model: (Constant), Systolic BP, Sex: numeric

13.2.10.2 Interpretation

During analysis using the “Stepwise” method, we included 5 independent variables (age, sex, systolic BP, and two dummy variables of religion – reli_1 and rali_2) in the model. The SPSS has provided the adjusted regression coefficients and two models (model 1 and 2) as shown in Table 13.17. Let us compare the outputs in Table 13.17 with those of Table 13.9, where we have used the “*Enter*” method. In Table 13.9, we can notice that SPSS retained all the independent variables in the model that were included, and only the “systolic BP” (p<0.001) and “sex” (p= 0.019) are found to be significantly associated with the dependent variable (diastolic BP). When we used the “*Stepwise*” method, SPSS has provided two models – model 1 and model 2 (Table 13.17). In model 1, there is only one independent variable (systolic BP) and in model 2, there are two independent variables (systolic BP and sex; others are automatically removed). We consider the last model (in this example, model 2) as the final model. The variables that have been removed (excluded) from each of the models are listed in Table 13.19. Table 13.18 shows the multiple R, R-square and adjusted R-square values of each model for comparison.

Occasionally, you may need to include certain variable(s) in the model for theoretical or practical reasons. In such a situation, after you derive the model with “Stepwise” method, add the additional variable(s) of your choice and re-run the model using the “Enter” method.

For automatic selection method, you can specify the inclusion (entry) and exclusion (removal) criteria of the variables. Usually, the inclusion and exclusion criteria, set as default in SPSS, are 0.05 and 0.10, respectively (Fig 13.8). You can, however, change the criteria based on your requirements. Finally, for model building, the researcher should decide the variables to be included in the final model based on theoretical understanding and empirical findings.

14

Logistic Regression

Logistic regression is the commonly used statistical method for health sciences data analysis. This method can be applied to analyze the data of cross-sectional, case-control or cohort studies. Logistic regression analysis is performed when the outcome variable is a categorical variable, either dichotomous (also called binary variable) (e.g., disease – present/absent), unordered polychotomous (e.g., type of food preferred – rice/bread/meat) or ordinal variable (severity of pain – mild/moderate/severe). The predictive (independent) variables can be either categorical or continuous. Like other multivariable analyses, the purposes of multivariable logistic regression analysis are to:

- Adjust the estimate of risk (Odds ratio) for a number of factors set in the model;
- Determine the relative contribution of factors to a single outcome;
- Predict the probability of an outcome for a number of independent variables in the model; and
- Assess interaction of multiple variables for the outcome.

The logistic regression analysis can be applied in several methods, depending on the type of outcome variable and study design. They are:

1. **Binary logistic regression**: This method is used when the dependent variable is a dichotomous (binary) categorical variable, such as disease (present/absent), vaccinated (yes/no), and outcome of the patient (died/survived). The binary logistic regression can be applied as:

   a) *Unconditional binary logistic regression*: This method is used when the dependent variable is a dichotomous categorical variable in an *unmatched* study design (e.g., unmatched case-control studies). Note that the term "*unconditional binary logistic regression*" is commonly and simply expressed as "*logistic regression*"; and

   b) *Conditional binary logistic regression*: This method is applied where the dependent variable is a dichotomous variable and the cases are *matched* with

with controls for one or more variables (e.g., matched case-control studies). The word “*binary*” is commonly omitted from the terminology and is simply expressed as “*conditional logistic regression*”.

2. **Multinominal logistic regression**: This method of data analysis is used when the outcome (dependent) variable is a nominal categorical variable with *more than two levels*, such as health seeking behaviour (did not seek treatment/ received treatment from village doctors/ received treatment from pharmacists), type of cancer (stomach cancer/ lung cancer/ skin cancer) and others.

3. **Ordinal logistic regression (proportional odds regression):** This method is used when the outcome variable is an ordinal categorical variable, like severity of pain (mild/ moderate/ severe) and stage of cancer (stage 1/ stage 2/ stage 3/ stage 4).

Mathematical concept of logistic regression model

In logistic regression analysis, Odds are transformed into natural log (ln) of Odds, i.e., lnOdds. Ln is the log to the base of e. To reverse the ln, we take the exponential (e^x) of the log value. When the Odds are transformed into lnOdds, it is called *logit transformation*. In logistic regression, lnOdds of the outcome variable are put on the Y-axis. The multivariable logistic regression model is given by the equation:

$$\ln\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Where, P denotes the probability of the outcome, β_0 is the intercept (constant) and β_i is the regression coefficient of the i^th^ variable (i= 1, 2, …, n) and X_i represents the values of the predictor (independent) variables in the model, $X_i = (X_1, X_2, \ldots X_n)$. The Exp[B], provided by SPSS during analysis, indicates the OR adjusted for all other variables in the model. Therefore, the regression coefficients (β) that we get in logistic regression analysis are the lnOdds, and the exponential of the regression coefficients (Exp[B] in SPSS) are the Odds Ratio (OR) for the categorical independent variables. If the independent variable is a continuous variable, the interpretation is different and is discussed in section 14.1.5. We can also calculate the probability (*p*) of an outcome, using the following formula:

$$p = \frac{Exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}{1 + Exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}$$

Or

$$p = \frac{1}{1 + Exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)]}$$

The detailed explanation of the model can be found in any standard Biostatistics book. In this chapter, we shall discuss the unconditional and conditional logistic regression methods. The multinominal logistic regression is discussed in chapter 15.

Assumptions

Logistic regression does not make any assumption concerning the distribution of predictor (independent) variables. However, it is sensitive to high correlation among the independent variables (multicollinearity). The outliers may also affect the results of logistic regression.

14.1 Binary logistic regression

Binary logistic regression analysis is appropriate when the outcome variable is a *dichotomous* categorical variable (e.g., disease present/absent) for the adjustment of multiple confounding factors or model (identify) the predictors of a dichotomous categorical outcome. For logistic regression analysis in SPSS, the dichotomous outcome variable is coded as "0= disease absent" and "1= disease present". *SPSS will consider the higher value to be the predicted outcome and the lower value as the comparison group*.

14.1.1 Unconditional logistic regression

The unconditional logistic regression is simply called logistic regression. We shall use the term "logistic regression" to mean the unconditional logistic regression throughout this chapter.

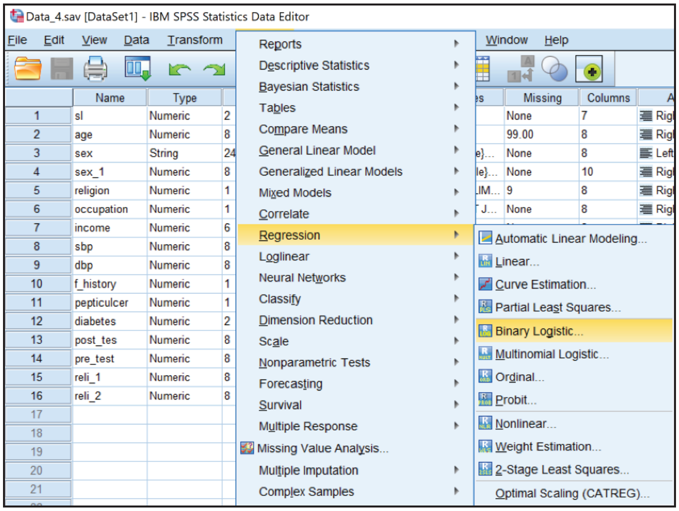
Suppose you have conducted an unmatched study (e.g., a cross-sectional or an unmatched case-control study) to identify the factors (or predictors) associated with diabetes (dependent variable). The independent variables/factors that you have considered for the analysis are: sex (variable name: sex_1), age, peptic ulcer (variable name: pepticulcer) and family history of diabetes (variable name: f_history). To perform the logistic regression analysis, recode diabetes as "0= diabetes absent" and "1= diabetes present" (if it is not coded like this). Similarly, it is better to recode the categorical independent (predictor) variables as "0 for no" (comparison group) and "1 for yes" (there is no problem, if you do not do this). Use the data file <**Data_4.sav**> for practice.

14.1.2 Commands

Analyze > Regression > Binary logistic > Put "diabetes" in the "Dependent" box > Put "age, sex_1, f_history and pepticulcer" in the "Covariates" box > Categorical >

Push “sex_1, f_history and pepticulcer” from “Covariates” box to “Categorical covariates” box > Select “sex_1” > Select “first” for “Reference category” under “Change contrast” [we are doing this because 0 (female) is our comparison group. Note that the default category for comparison is the last category] > Click on “Change” > (Do the same thing for all the variables in “Categorical covariates” box) > Continue > Options > Select “Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals, Correlations of estimates, and CI for exp(B)” > Continue > OK (Figs 14.1 to 14.3)

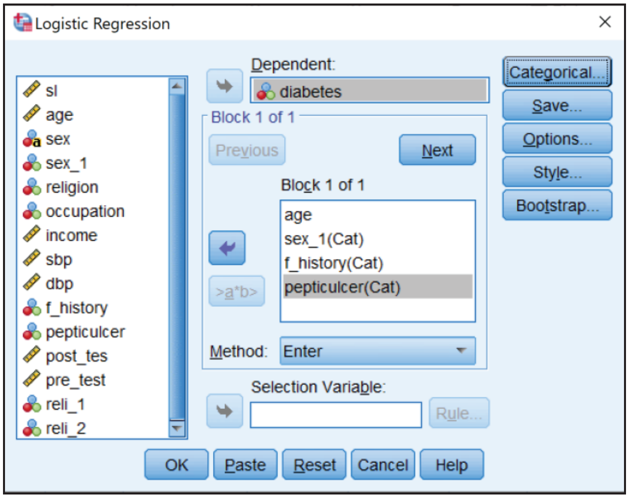
Figure 14.1



Data_4.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type		s	Missing	Columns	A
1	sl	Numeric	2		None	7	Rig
2	age	Numeric	8		99.00	8	Rig
3	sex	String	24	e}...	None	8	Left
4	sex_1	Numeric	8	le}...	None	10	Rig
5	religion	Numeric	1	IM...	9	8	Rig
6	occupation	Numeric	1	J...	None	8	Rig
7	income	Numeric	6				
8	sbp	Numeric	8				
9	dbp	Numeric	8				
10	f_history	Numeric	1				
11	pepticulcer	Numeric	1				
12	diabetes	Numeric	2				
13	post_tes	Numeric	8				
14	pre_test	Numeric	8				
15	reli_1	Numeric	8				
16	reli_2	Numeric	8				
17							
18							
19							
20							
21							

Figure 14.2



Logistic Regression

Dependent: diabetes

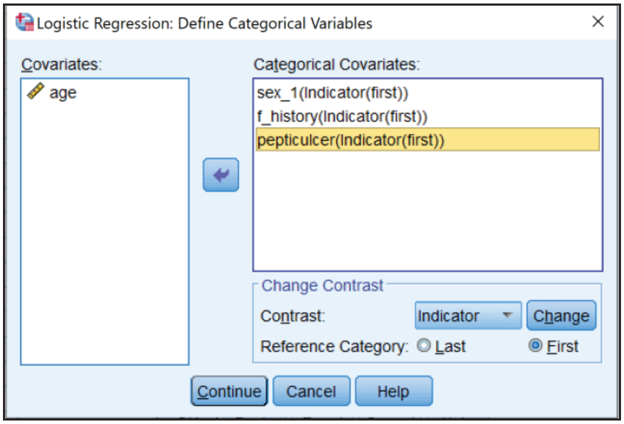
Block 1 of 1

age
sex_1(Cat)
f_history(Cat)
pepticulcer(Cat)

Method: Enter

Selection Variable:

Figure 14.3



Logistic Regression: Define Categorical Variables

Covariates:

age

Categorical Covariates:

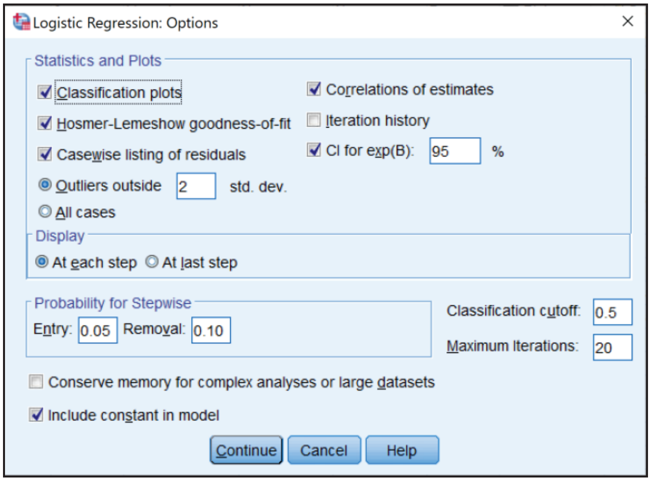
sex_1(Indicator(first))
f_history(Indicator(first))
pepticulcer(Indicator(first))

Change Contrast

Contrast: Indicator

Reference Category: ○ Last ● First

Figure 14.4 Option template for logistic regression



Logistic Regression: Options

Statistics and Plots

☑ Classification plots
☑ Hosmer-Lemeshow goodness-of-fit
☑ Casewise listing of residuals
● Outliers outside 2 std. dev.
○ All cases
☑ Correlations of estimates
□ Iteration history
☑ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10

Classification cutoff: 0.5

Maximum Iterations: 20

□ Conserve memory for complex analyses or large datasets

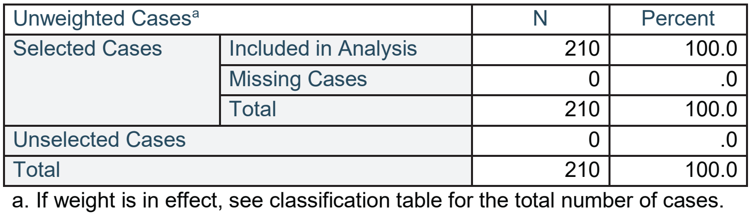
☑ Include constant in model

14.1.3 Outputs

SPSS provides many tables while doing the logistic regression analysis. Only the useful tables are discussed here. After the basic tables (Tables 14.1 to 14.3), the outputs of logistic regression are provided under Block 0 and Block 1.

A. Basic tables

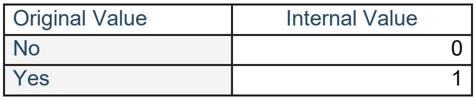
Table 14.1 Case Processing Summary



Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	210	100.0
	Missing Cases	0	.0
	Total	210	100.0
Unselected Cases		0	.0
Total		210	100.0

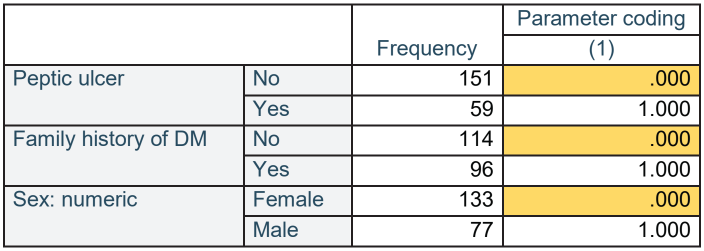
a. If weight is in effect, see classification table for the total number of cases.

Table 14.2 Dependent Variable Encoding



Original Value	Internal Value
No	0
Yes	1

Table 14.3 Categorical Variables Codings



		Frequency	Parameter coding
			(1)
Peptic ulcer	No	151	.000
	Yes	59	1.000
Family history of DM	No	114	.000
	Yes	96	1.000
Sex: numeric	Female	133	.000
	Male	77	1.000

14.1.4 Interpretation: Basic tables

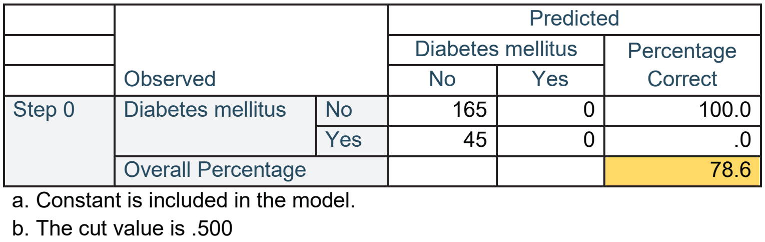
Table 14.1 (case processing summary) shows that the analysis includes all the 210 subjects and there is no missing value. If there are any missing data, this table will show it. Note that the subjects are excluded from the analysis, if there are any missing data.

Table 14.2 (dependent variable encoding) tells us which category of the dependent variable (diabetes) is the predicted outcome. The higher value is the predicted outcome. Here, the higher value is 1 (have diabetes) and is the predicted outcome for the dependent variable.

Table 14.3 (categorical variables codings) indicates the comparison groups of the independent (explanatory) variables. Here, the lower value is the comparison group. For example, for family history of diabetes, "No" is coded as 0 (.000), while "Yes" is coded as 1 (1.000). This means that persons who do not have family history of diabetes are in the comparison group (reference category). Similarly, not having peptic ulcer (.000) and being female (.000) are the comparison groups for peptic ulcer and sex, respectively.

B. Outputs under Block 0

Table 14.4 Classification Table[a,b]



			Predicted		
			Diabetes mellitus		Percentage Correct
	Observed		No	Yes	
Step 0	Diabetes mellitus	No	165	0	100.0
		Yes	45	0	.0
	Overall Percentage				78.6

a. Constant is included in the model.

b. The cut value is .500

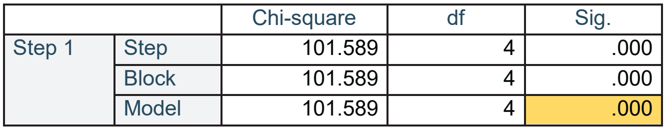
14.1.5 Interpretation: Outputs under Block 0

Analysis of data without any independent variable in the model is provided under Block 0. The results indicate the baseline information that can be compared with the results when independent variables are included in the model (provided under Block 1).

Look at the classification table (Table 14.4). The table indicates the *overall percentage* of correctly classified cases (78.6%). We will see whether this value has increased with the introduction of independent variables in the model under Block 1 (given in Table 14.8). If the value remains the same, it means that the independent variables in the model do not have any influence/contribution in predicting diabetes (dependent variable). In our example, the overall percentage has increased to 90.5% after inclusion of the independent variables in the model as shown in Table 14.8 under Block 1 compared to the value under Block 0 (78.6%). This means that addition of independent variables improved the ability of the model to predict the dependent variable.

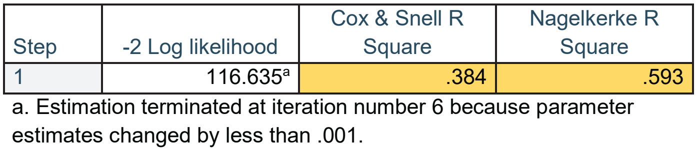
C. Outputs under Block 1

Table 14.5 Omnibus Tests of Model Coefficients



		Chi-square	df	Sig.
Step 1	Step	101.589	4	.000
	Block	101.589	4	.000
	Model	101.589	4	.000

Table 14.6 Model Summary



Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	116.635[a]	.384	.593

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Table 14.7 Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.663	8	.066

Table 14.8 Classification Table[a]

			Predicted		
			Diabetes mellitus		
	Observed		No	Yes	Percentage Correct
Step 1	Diabetes mellitus	No	159	6	96.4
		Yes	14	31	68.9
	Overall Percentage				90.5

a. The cut value is .500

Table 14.9 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age in years	.230	.039	34.660	1	.000	1.259	1.166	1.359
	Sex: numeric(1)	1.587	.528	9.031	1	.003	4.891	1.737	13.774
	Family history of DM(1)	1.119	.517	4.688	1	.030	3.062	1.112	8.430
	Peptic ulcer(1)	1.782	.487	13.363	1	.000	5.942	2.285	15.448
	Constant	-10.684	1.526	49.006	1	.000	.000		

a. Variable(s) entered on step 1: age in years, Sex: numeric, Family history of DM, Peptic ulcer.

Table 14.10 Correlation Matrix

		Constant	age in years	Sex: numeric(1)	Family history of DM(1)	Peptic ulcer(1)
Step 1	Constant	1.000	-.932	-.385	-.302	-.322
	age in years	-.932	1.000	.161	.054	.156
	Sex: numeric(1)	-.385	.161	1.000	.388	.155
	Family history of DM(1)	-.302	.054	.388	1.000	.099
	Peptic ulcer(1)	-.322	.156	.155	.099	1.000

Table 14.11 Casewise List[b]

		Observed			Temporary Variable		
Case	Selected Status[a]	Diabetes mellitus	Predicted	Predicted Group	Resid	ZResid	SResid
11	S	Y**	.006	N	.994	13.213	3.220
25	S	Y**	.026	N	.974	6.177	2.728
38	S	Y**	.052	N	.948	4.248	2.443
41	S	Y**	.062	N	.938	3.883	2.389
62	S	Y**	.099	N	.901	3.009	2.167
124	S	Y**	.043	N	.957	4.692	2.521
137	S	N**	.890	Y	-.890	-2.839	-2.140

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

14.1.6 Interpretation: Outputs under Block 1

Omnibus tests of Model Coefficients (Table 14.5): This table indicates whether the overall performance of the model is improved when independent variables are included in the model compared to the model without any independent variables (given under Block 0). We want this test to be significant (p-value < 0.05). In this example, the p-value of the Omnibus test is 0.000, which indicates that the proposed model is better than the model without any predictor (independent) variables.

Model summary table (Table 14.6): This table indicates usefulness of the model. The Cox & Snell R-square and Nagelkerke R-square (called pseudo R-square) values provide an indication about the amount of variation in the outcome variable that can be explained by the independent variables in the model. In this example, the values of the pseudo R-square are 0.384 (Cox & Snell R-square) and 0.593 (Nagelkerke R-square), respectively. This means that between 38.4% and 59.3% variation in the outcome variable can be explained by the independent variables set in the model. *This information is not needed if the objective of the analysis is to adjust for the Odds Ratio*.

Hosmer-Lemeshow goodness-of-fit test (Table 14.7): When the intention of analysis is prediction, i.e., to identify the predictors to predict the outcome, then the question is "How good is the model for prediction"? This is judged based on the Hosmer-Lemeshow goodness-of-fit test, and positive and negative predictive values, given in the classification table (Table 14.8).

The Hosmer-Lemeshow test indicates how well the observed and predicted values fit with each other (i.e., observed and predicted probabilities match with each other). The null hypothesis is "the model fits" and the p-value is expected to be >0.05 (non-significant). If the p-value is not significant, it means that the model is a good fit for prediction (i.e., the observed and predicted values are close together). In this example, the p-value is 0.066, indicating that the model is useful for prediction of the outcome variable. If the test is significant (p<0.05), then the model is not good to predict the outcome variable by the independent variables in the model. *Note that this information is not needed if the objective of doing logistic regression is to adjust for the confounding factors*.

Classification table (Table 14.8): This table indicates how well the model is able to predict the correct category of the dependent variable (have or do not have the disease). This table shows that the overall accuracy of this model to predict diabetes (with a predicted probability of 0.5 or greater) is 90.5%. This table also shows the *Sensitivity* and *Specificity* of the model as 68.9% [31 ÷ (14+31)] and 96.4% [159 ÷ (159+6)], respectively. Positive and negative predictive values can also be calculated from the table, which are

83.8% [31 ÷ (31+6)] and 91.9% (159 ÷ [(159+14)], respectively. Interpretation of the findings of this table is a little bit complicated and needs further explanation, especially to explain the sensitivity, specificity, and positive and negative predictive values[8].

However, the information that we need to check is the *overall percentage*. Compare this value with the value under the Block 0 outputs. We expect this value (overall percentage) to be increased, otherwise adding independent variables in the model does not have any impact on prediction. We can see that the overall percentage of the model to correctly classify cases is 90.5% under Block 1 (Table 14.8). This value, compared to the value (78.6%; Table 14.4) that we have seen under Block 0, has improved. This means that adding independent variables in the model improved the ability of the model to predict the dependent variable. *This information is needed, if the intention of this analysis is prediction. If the objective is adjustment for confounding factors, we can ignore this information.*

Variables in the equation (Table 14.9): *This is the most important table to look at*. This table shows the results of logistic regression analysis. This table indicates how much each of the independent variables contributes to predicting/explaining the outcome variable. This table also indicates the adjusted Odds Ratio (OR) and its 95% confidence interval (CI). The B values (column 3) indicate the logistic regression coefficients of the variables in the model. These values are used to calculate the probability of an individual to have the outcome. The positive values indicate the likelihood for the outcome, while the negative values indicate the less likelihood for the outcome. The exponential of B [Exp(B)] is the adjusted OR for the categorical variables in the model.

Let us see how to interpret the results. There are 4 independent (explanatory) variables in the model – age (as a continuous variable), sex, family history of diabetes and peptic ulcer. The table shows the adjusted OR [Exp(B)], 95% CI for the adjusted OR and p-value (Sig.). The adjusted OR for sex is 4.891 (95% CI: 1.737 to 13.774), which is statistically significant (p=0.003). Here, our comparison group is female (see Table 14.3). This indicates that males are 4.9 times more likely to have diabetes compared to females after adjusting (or controlling) for age, family history of diabetes and peptic ulcer. Similarly, persons who have the family history of diabetes are 3.1 times more likely (OR: 3.06; 95% CI: 1.11 to 8.43; p=0.03) to have diabetes compared to those who do not have a family history of diabetes after adjusting for age, sex, and peptic ulcer. Interpretation of Exp(B) for age is a little bit different since the variable was entered as a continuous variable. In our example, the Exp(B) for age is 1.259. This means that the odds of having diabetes will increase by 25.9% [Exp(B) – 1; i.e., 1.259 – 1] (95% CI: 16.6 to 35.9) with each year increase in age, which is statistically significant (p<0.001).

If we want to know which variable contributed most to the model, then we have to look at the Wald statistics. Higher the value (of Wald), the greater is the importance.

Therefore, “age” is the most important variable contributing to the model since it has the highest Wald value (34.6).

Checking for multicollinearity (Table 14.10): It is important to check for multicollinearity of the independent variables in the model. If there is multicollinearity, the model becomes dubious. Multicollinearity is checked in the correlation matrix table (Table 14.10). This table shows the correlations (correlation coefficients or “r” values) between the independent variables. If there is multicollinearity, “r” values will be high (greater than 0.5). If we look at the correlation matrix table (Table14.10), none of the values is greater than 0.5 except for the correlation between age and constant, which is -0.932.

Now, look at Table 14.9 (variables in the equation). If multicollinearity is present (and affects the model), the magnitude of the SEs (standard errors) will be high or low (greater than 5.0 or less than 0.001). The existence of multicollinearity means that the model is not statistically stable. To solve the problem (in general), look at the SE and omit the variable(s) with large (or small) SE, until the magnitude of the SEs hovers *between 0.001 and 5.0*.

If there is a high correlation between the constant and any of the predictor variables, you can omit the constant from the model. In our example, there is a high correlation (-.932) between age and constant. However, it did not affect the results as none of the SEs is >5.0 or <0.001. Therefore, we do not need to do anything. If it affects the results, just omit (deselect) the constant from the model *(deselect “Include constant in model” located at the bottom of the option template)* from “Options template” during analysis as shown in Figure 14.4.

Casewise list (Table 14.11): Table 14.11 provides information about the cases for which the model does not fit well. Look at the ZResid values (last column of the table). The values above 2.5 are the outliers and do not fit well in the model. The case numbers are shown in column 1. If present (cases that do not fit the model well), all these cases need to be examined closely. Under the “Predicted Group” column, you may see “Y (means yes)” or “N (means no)”. If it is “Y”, the model predicts that the case (case no. 137, in our example) should have diabetes, but in reality (in the data) the subject does not have diabetes (see the observed column where it is “N”). Similarly, if it is “N” under the “Predicted Group”, the model predicts that the case should not have diabetes, but in the data the subject has diabetes.

14.1.7 ROC curve

We can construct the receiver-operating characteristic (ROC) curve to assess the model discrimination, which is an indication of accuracy of the logistic regression model. Discrimination is defined as the ability of the model to distinguish between those who have the outcome (e.g., disease) and those who do not have the outcome. Discrimination is evaluated by using the ROC curve analysis. In ROC curve analysis, the area under the curve (termed as c-statistic) is measured.

The area under the ROC curve ranges from 0 to 1. A value of 0.5 indicates that the model is useless. Values between 0.7 and 0.8 are considered as acceptable discrimination, values between 0.8 and 0.9 indicate excellent discrimination, and values ≥0.9 indicate outstanding discrimination.

Our data shows that the area under the ROC curve is 0.914 (95% CI: 0.86 to 0.96; p=0.000) (Table 14.12). Since the value is >0.9, it indicates an excellent model for prediction. To generate the ROC curve, use the following commands.

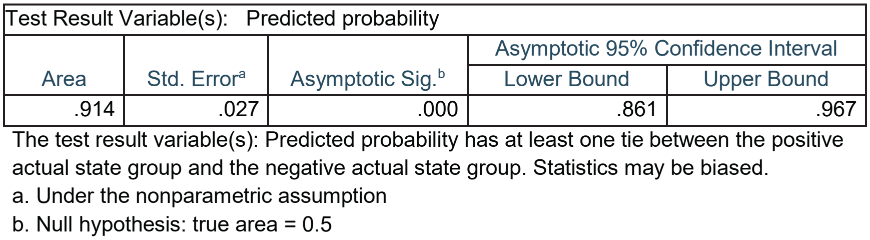
> Analyze > Regression > Binary logistic > Put "diabetes" in the "Dependent" box > Put "age, sex_1, f_history and pepticulcer" in the "Covariate" box > Categorical > Push "sex_1, f_history and pepticulcer" from "Covariates" box to "Categorical covariates" box > Select "sex_1" > Select "first" for "Reference category" under "Change contrast" [we are doing this because 0 (female) is our comparison group. Note that the default category for comparison is the last category] > Click on "Change" > (Do the same thing for all the variables in "Categorical covariates" box) > Continue > Options > Select "Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals, Correlations of estimates, and CI for exp(B)" > Continue > Click on "**Save**" > Select "Probabilities" under "Predicted values" > Continue > OK

This will generate a new variable, PRE_1 (predicted probability) (look at the bottom of the SPSS variable view). Now, to get the ROC curve, use the following commands.

> Analyze > ROC curve > Select "**PRE_1**" for the "Test variable" box and "diabetes" for the "State variable" box and write "1" in "Value of state variable" box (since code 1 indicates individuals with diabetes) > Select "ROC curve" and "Standard error and Confidence interval" under "Display" > OK

With these commands, SPSS will generate Table 14.12 and the ROC curve (Fig 14.5).

Table 14.12 Area Under the Curve



Test Result Variable(s): Predicted probability				
			Asymptotic 95% Confidence Interval	
Area	Std. Error[a]	Asymptotic Sig.[b]	Lower Bound	Upper Bound
.914	.027	.000	.861	.967

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

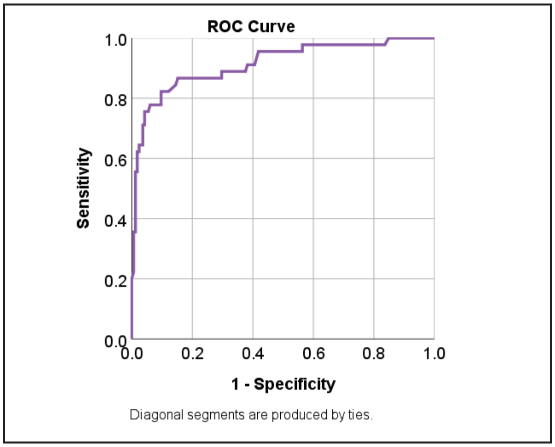
a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

14.1.8 Sample size for logistic regression

Sample size is always a concern for analysis of data. The sample size needed for logistic regression depends on the effect size you are trying to demonstrate and the variability of the data. It is always better to calculate the sample size during the design phase of the study. However, a rule-of-thumb for planning a logistic regression analysis is that for every independent variable in the model you need to have at least 10 outcomes (some authors recommend a minimum of 15-25 cases for each independent variable[4, 14]).

Figure 14.5 ROC curve



14.1.9 Variable selection for a model

We have discussed different methods of variable selection for a model in the previous chapter in detail (chapter 13). Like other multivariable analyses, independent variables to be selected for logistic regression should include the risk factors of interest and potential confounders, while avoiding variables with lots of missing values.

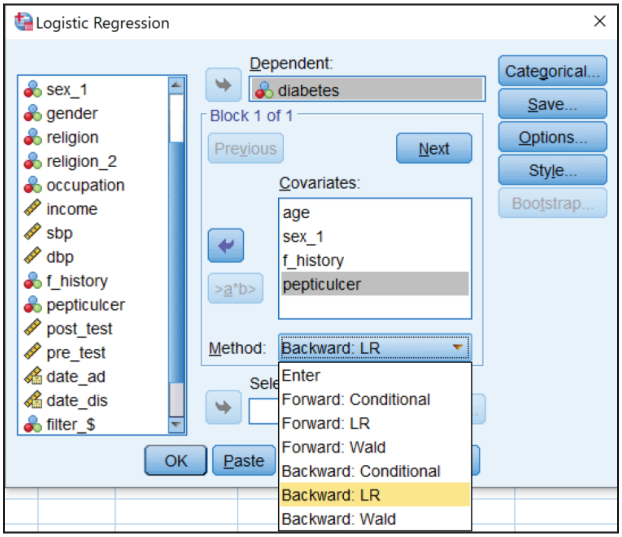
So far in this chapter, we have used the "Enter" method for logistic regression analysis. The "Enter" method uses all the independent variables in the model included by the researcher. We can use the automatic selection method for analysis as well. For logistic regression, the commonly used method for automatic selection of variables is the "*Backward LR*" method. However, if there is multicollinearity, you may select the "*Forward LR*" method during analysis.

14.1.9.1 Commands for automatic selection of independent variables (use the data file <Data_3.sav>):

> Analyze > Regression > Binary logistic > Put "diabetes" in the "Dependent" box > Put "age, sex_1, f_history and pepticulcer" in the "Covariates" box > **Select "Backward LR" from "Method"** (Fig 14.6) > Categorical > Push "sex_1, f_history and pepticulcer" from "Covariates" box to "Categorical covariates" box > Select "sex_1" > Select "first" for "Reference category" under "Change contrast" > Click on "Change" > (Do the same thing for all the variables in "Categorical covariates" box) > Continue > Options > Select "Hosmer-Lemeshow goodness-of-fit and CI for exp(B)" > Continue > OK

The SPSS will produce the table "Variables in the equation" (Table 14.13) along with others (not shown here as they are not relevant). We can see that the analysis is completed in 4 steps (Step 1 to 4; the first column of the table). In the first step, all the independent variables are in the model. Gradually, SPSS has removed variables that are not significantly associated with the outcome. Finally, SPSS has provided the final model (Step 4) with a single variable (sex) in it, which is significantly associated with the outcome. If the "Enter" method is used, SPSS will provide only step 1 (Table 14.14).

Figure 14.6



Logistic Regression

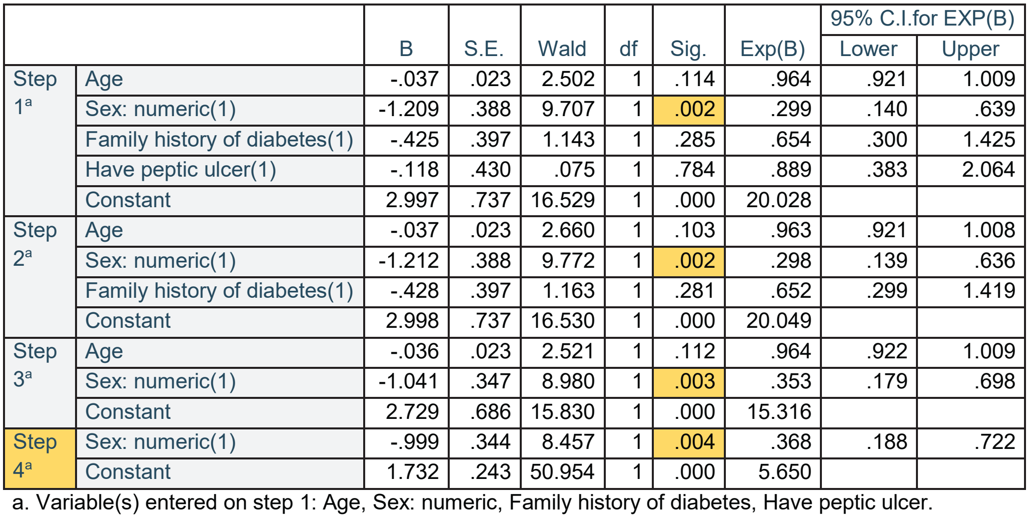
Dependent: diabetes

Block 1 of 1

Covariates: age, sex_1, f_history, pepticulcer

Method: Backward: LR

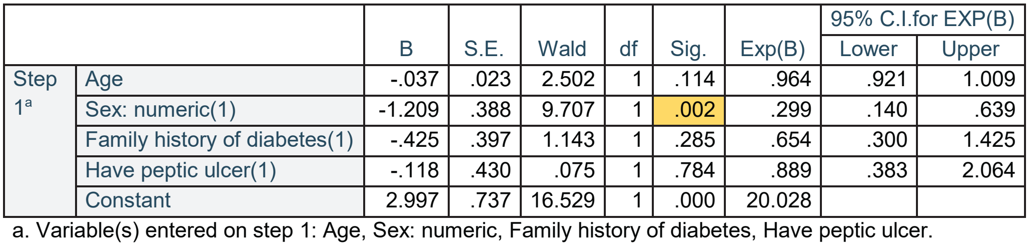
Table 14.13 Variables in the Equation: Backward LR method



		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	-.037	.023	2.502	1	.114	.964	.921	1.009
	Sex: numeric(1)	-1.209	.388	9.707	1	.002	.299	.140	.639
	Family history of diabetes(1)	-.425	.397	1.143	1	.285	.654	.300	1.425
	Have peptic ulcer(1)	-.118	.430	.075	1	.784	.889	.383	2.064
	Constant	2.997	.737	16.529	1	.000	20.028		
Step 2[a]	Age	-.037	.023	2.660	1	.103	.963	.921	1.008
	Sex: numeric(1)	-1.212	.388	9.772	1	.002	.298	.139	.636
	Family history of diabetes(1)	-.428	.397	1.163	1	.281	.652	.299	1.419
	Constant	2.998	.737	16.530	1	.000	20.049		
Step 3[a]	Age	-.036	.023	2.521	1	.112	.964	.922	1.009
	Sex: numeric(1)	-1.041	.347	8.980	1	.003	.353	.179	.698
	Constant	2.729	.686	15.830	1	.000	15.316		
Step 4[a]	Sex: numeric(1)	-.999	.344	8.457	1	.004	.368	.188	.722
	Constant	1.732	.243	50.954	1	.000	5.650		

a. Variable(s) entered on step 1: Age, Sex: numeric, Family history of diabetes, Have peptic ulcer.

Table 14.14 Variables in the Equation: Enter method



		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	-.037	.023	2.502	1	.114	.964	.921	1.009
	Sex: numeric(1)	-1.209	.388	9.707	1	.002	.299	.140	.639
	Family history of diabetes(1)	-.425	.397	1.143	1	.285	.654	.300	1.425
	Have peptic ulcer(1)	-.118	.430	.075	1	.784	.889	.383	2.064
	Constant	2.997	.737	16.529	1	.000	20.028		

a. Variable(s) entered on step 1: Age, Sex: numeric, Family history of diabetes, Have peptic ulcer.

The inclusion (entry) and exclusion (removal) criteria, set as default in SPSS, are 0.05 and 0.10, respectively. As discussed in section 13, you can change the "Entry" and "Removal" criteria from the "Option" dialogue box (under the "Probability for Stepwise") (Fig 14.4). Finally, for model building, you should decide the variables to be included in the final model based on theoretical understanding and empirical findings.

14.1.10 Incorporating interaction terms in the model

Interaction and confounding are not the same. Interaction is sometimes called effect modification. Interaction indicates that the presence of a third variable (effect modifier) influences the relationship between an independent and dependent variable. In other words, when a risk factor's effect on outcome is changed by the value (category) of a third variable (interaction variable), interaction is said to be present.

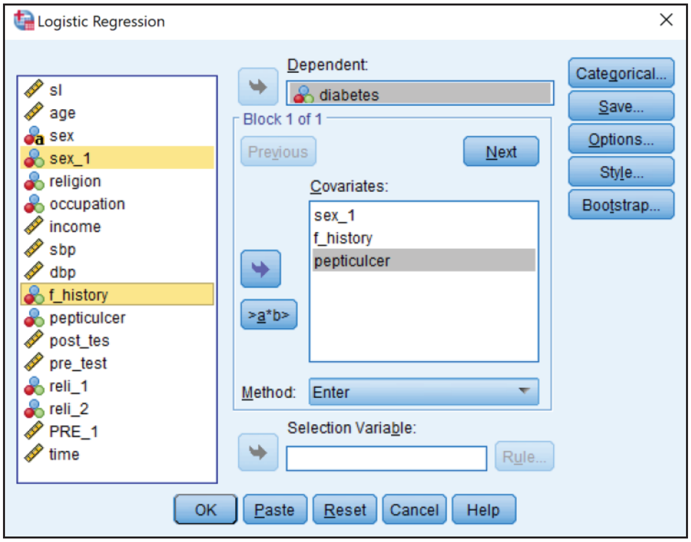
You can add interaction terms in logistic regression analysis with other variables for adjustment. Suppose that we are interested in determining the effect of sex on diabetes.

We are also keen to know if there is any interaction between sex (variable name is sex_1) and family history of diabetes (variable name is "f_history"). In such a situation, we shall have to include the interaction term "sex_1*f_history" in the model. Note that the variables for which we are looking for interaction, must be included in the model independently. To add interaction term in the model (use the data file <**Data_4.sav**>), use the following commands.

> Analyze > Regression > Binary logistic > Put "diabetes" in the "Dependent" box > Put "sex_1, f_history and pepticulcer" in the "Covariates" box > Click on "sex_1" and then on "f_history" pressing the "Ctrl" key on the keyboard (you will see that ">**a*b**>" button is highlighted (Fig 14.7) > Click on ">**a*b**>" (you will see "f_history(Cat)*sex_1(Cat)" in the "Block 1 of 1" box) > Categorical > Push "sex_1, f_history and pepticulcer" from "Covariates" box to "Categorical covariates" box > Select "sex_1, f_history and pepticulcer" all together pressing the "Shift" key on the keyboard > Select "first" for "Reference category" under "Change contrast" > Click on "Change" > Continue > Options > CI for exp(B)" > Continue > OK

The above commands will generate Table 14.15 (variables in the equation) along with the other tables. The interaction variable included in the model is shown as "Family history of DM(1) by Sex: numeric(1)". The table shows that the p-value for interaction is 0.081 (>0.05), indicating that there is no interaction between sex and family history of diabetes for the outcome (i.e., effect of sex on diabetes is not dependent on the family history of diabetes).

Figure 14.7



Dependent: diabetes

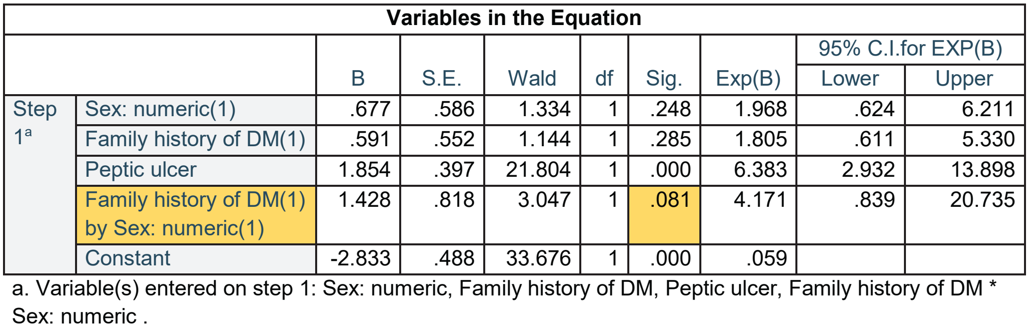
Block 1 of 1

Covariates:
sex_1
f_history
pepticulcer

Method: Enter

Selection Variable:

Table 14.15 Variables in the Equation with interaction between sex and family history of diabetes



Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Sex: numeric(1)	.677	.586	1.334	1	.248	1.968	.624	6.211
	Family history of DM(1)	.591	.552	1.144	1	.285	1.805	.611	5.330
	Peptic ulcer	1.854	.397	21.804	1	.000	6.383	2.932	13.898
	Family history of DM(1) by Sex: numeric(1)	1.428	.818	3.047	1	.081	4.171	.839	20.735
	Constant	-2.833	.488	33.676	1	.000	.059		

a. Variable(s) entered on step 1: Sex: numeric, Family history of DM, Peptic ulcer, Family history of DM * Sex: numeric .

14.2 Conditional logistic regression

In section 14.1 we have discussed the unconditional logistic regression for an unmatched design. The conditional logistic regression is done when the cases and controls are matched (e.g., a matched case-control design) for one or more variables.

Suppose we have conducted a matched case-control study (cases and controls are matched for gender) to identify the risk factors for death due to COVID infection. The outcome of our matched case-control study is death due to COVID (case) and the control is who survived the COVID infection. Each case is matched with a person who survived by gender. Use the data file <**Data_Ca-Co_matched.sav**> for practice (the variable "mID" is the matching ID number). In this dataset, the variable "death" indicates whether the person survived/died due to COVID and is coded as 0= survived (control) and 1= died (case).

Conditional logistic regression analysis in SPSS can be done either by the multi-nominal logistic regression method or Cox regression method. It is easier to perform the analysis using the Cox regression method and is discussed below.

14.2.1 Commands

14.2.1.1 Commands for generating the time variable

Before we perform the Cox regression, we need to generate a "time" variable for the use during analysis. The "time" variable should be coded as 1 if it is a case (death) and 2 if it is a control (survived). Use the following commands, to generate the "time" variable.

> Transform > Compute Variable > Type "time" in "Target Variable" box > Click in the box under "Numeric Expression" > Write "1" using the number pad or keyboard > OK (also see section 6.3)
>
> Again,

Transform > Compute Variable > Click in the box under "Numeric Expression" > Delete 1 > Type 2 using the "number pad" (or the keyboard) > Click "If (optional case selection condition)" > Select "Include if case satisfies condition" > Select "death" and push it into the empty box > Click "equal to sign (=)" and then write "0" using the "number pad" (always use the number pad) > Continue > OK > SPSS will provide the message "Change existing variable?" > Click "OK"

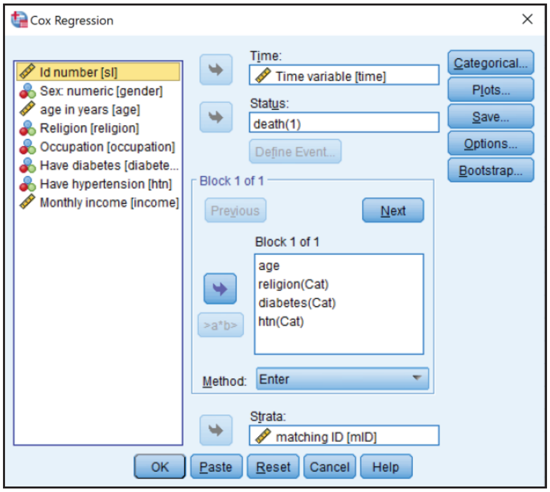
This will generate the "time" variable (the last variable) with 1 for case and 2 for control.

14.2.1.2 Commands for conditional logistic regression using Cox regression method

Suppose we want to evaluate whether age, religion, diabetes and hypertension (variable name is "htn") are the risk factors for COVID deaths in this analysis. Use the following commands to analyze the data.

Analyze > Survival > Cox Regression > Push the variable "time" into the "Time" box > Push "death" into the "Status" box > Click on "Define event" > In "Single value" box write 1 (since 1 is the code no. of the event/died) > Continue > Push the variables "age, religion, diabetes and htn" into the "Covariate" box > Select the variable "mID" and push it into the "Status" box > Click on "Categorical" > Push "religion, diabetes and htn" into the "Categorical covariates" box > Select "all the variables" in the "Categorical covariates" box > Select "First" from "Reference category" under "Change contrast" (we are doing this to specify the comparison group) > Change > Continue > Click on "Options" > Select "CI for Exp(B) and Correlation of estimates" > Continue > OK (Fig 14.8)

Figure 14.8



Cox Regression

Id number [sl]
Sex: numeric [gender]
age in years [age]
Religion [religion]
Occupation [occupation]
Have diabetes [diabete...
Have hypertension [htn]
Monthly income [income]

Time: Time variable [time]

Status: death(1)

Block 1 of 1

age
religion(Cat)
diabetes(Cat)
htn(Cat)

Method: Enter

Strata: matching ID [mID]

14.2.2 Outputs

SPSS will produce several tables. Only the relevant tables are provided below.

Tab 14.16 Coding scheme of categorical Variables in the model

Categorical Variable Codings[a,c,d]				
		Frequency	(1)	(2)
Religion[b]	1=MUSLIM	123	0	0
	2=HINDU	53	1	0
	3=Christian	24	0	1
Have diabetes[b]	0=No	107	0	
	1=Yes	93	1	
Have hypertension[b]	0=No	143	0	
	1=Yes	57	1	

a. Category variable: Religion (religion)
b. Indicator Parameter Coding
c. Category variable: Have diabetes (diabetes)
d. Category variable: Have hypertension (htn)

Table 14.17 Logistic regression analysis: adjusted odds ratio

Variables in the Equation								
							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
age in years	.074	.026	8.235	1	.004	1.077	1.024	1.133
Religion			6.218	2	.045			
Religion(1)	-.643	.452	2.020	1	.155	.526	.217	1.276
Religion(2)	-2.566	1.104	5.398	1	.020	.077	.009	.670
Have diabetes	.781	.419	3.465	1	.063	2.183	.960	4.967
Have hypertension	.944	.451	4.369	1	.037	2.569	1.061	6.225

14.2.3 Interpretation

The interpretations are similar to the unconditional logistic regression as discussed earlier. The main table of our interest is the "variables in the equation" table (Table 14.17) that shows the OR [Exp(B)] after adjusting for the variables in the model. The table shows that age (p=0.004), religion (Christians compared to Muslims) (p=0.020) and having hypertension (compared to no hypertension) (p=0.037) are the factors significantly associated with death due to COVID. Data indicates that Christians are less likely to die compared to Muslims (adjusted OR: 0.077; 95% CI: 0.009-0.670; p=0.020), while those who have hypertension are more than 2.5 times more likely to die compared to those who do not have hypertension (adjusted OR: 2.569; 95% CI: 1.06-6.22; p=0.037). Interpretation of Exp(B) for age is different since the variable was entered as a continuous variable (see section 14.1.5). In this example, the Exp(B) for age (1.077) indicates that the odds of dying increases by 7.7% (95% CI: 2.4-13.3%) with each year increase in age, which is statistically significant (p=0.004).

15

Multinominal Logistic Regression

Multinominal logistic regression is an extension of the binary logistic regression where the dependent variable is a nominal categorical variable (e.g., health seeking behaviour, type of cancer, cause of death) with more than two levels. Use the data file <**Data_5 multinominal.sav**> for practice.

For instance, a researcher has collected data to identify the factors associated with health seeking behaviour (variable name is "behaviour") for diarrhoea among children. Here the dependent variable is "health seeking behaviour", which has three levels,

1) Did not seek treatment (coded as 1);
2) Received treatment from the village doctor (coded as 2); and
3) Received treatment from the pharmacist (coded as 3).

The independent (explanatory) variables included in the analysis are: maternal age (variable name is "age" and included as a continuous variable), severity of diarrhoea (variable name is "s_diarrhoea" and coded as 1= severe; 2= not severe) and religion (1= Muslim; 2= others).

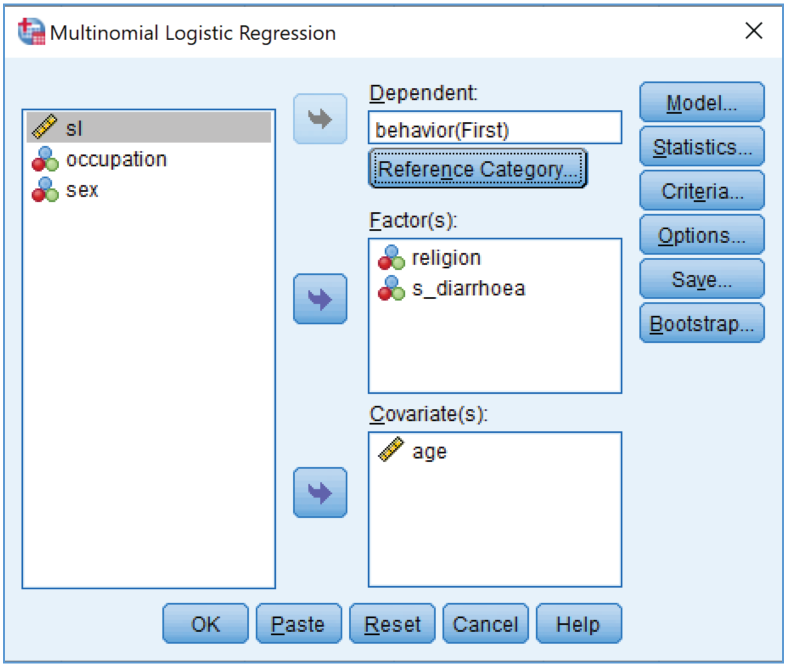
For multinominal logistic regression analysis, it is necessary to select a reference group/category of the dependent variable for comparison. The reference group selected for this analysis is "did not seek treatment". The analysis will, therefore, provide the estimates for the categories "received treatment from the village doctor" compared to "did not seek treatment", and "received treatment from the pharmacist" compared to "did not seek treatment". In this example, we have included two categorical variables (religion and severity of diarrhoea) and one quantitative variable (age) as explanatory variables in the model.

15.1 Commands

Analyze > Regression > Multinominal Logistic > Select the variable "behavior" and push it into the "Dependent" box > Click on "Reference Category" > Select "First category” under “Reference Category” > Continue > Select the variables “religion” and “s_diarrhoea” and push them into the “Factor(s)” box > Select “age” and push it into the “Covariate(s)” box (Fig 15.1) > Click on “Model” > Select “Main effects” (usually the default) > Continue > Click on “Statistics” > Select “Pseudo R-square, Step summary, Model fitting information, Classification table, and Goodness-of-fit” under “Model” > Select “Estimates and Likelihood ratio tests” under “Parameters” > Select “Covariate patterns defined by factors and covariates” under “Define Subpopulations” (Fig 15.2) > Continue > OK

Note: The quantitative variable(s) must be entered into the Covariate(s) box, while the categorical variables are entered into the Factor(s) box. The last category/value of the categorical independent variables is the comparison group by default in SPSS. In our example, religion is coded as 1= Muslim and 2= others. Therefore, the comparison group is the other religions, and the estimate will be for Muslims compared to other religions. For model selection (Model option), selection of “Main effect” will include all the variables specified in the analysis without any interaction term. Selection of “Full factorial” will provide the main effects with all the possible interactions, while the “Custom/-Stepwise” option provides freedom to set up relevant main effects and interaction terms using the “Forced Entry” option or to perform the Stepwise analysis.

Figure 15.1



Multinomial Logistic Regression

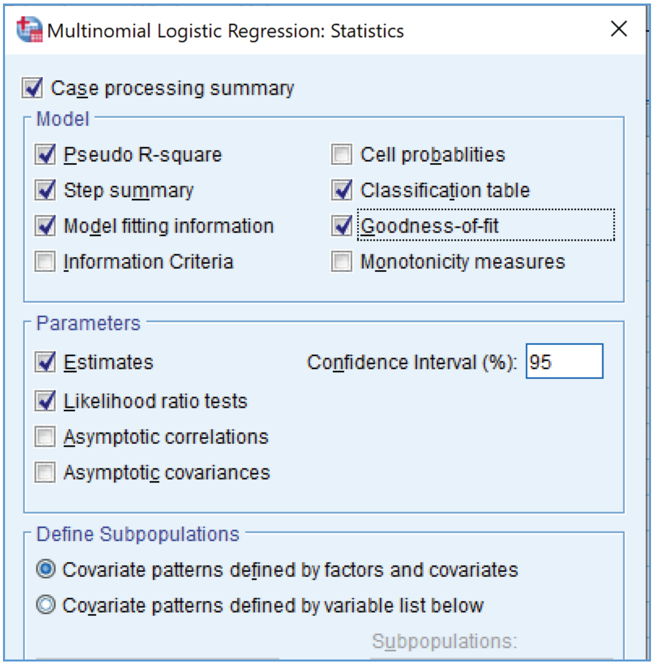
sl
occupation
sex

Dependent: behavior(First)

Factor(s):
religion
s_diarrhoea

Covariate(s):
age

Figure 15.2



Multinomial Logistic Regression: Statistics

☑ Case processing summary

Model

☑ Pseudo R-square
☑ Step summary
☑ Model fitting information
□ Information Criteria
□ Cell probablities
☑ Classification table
☑ Goodness-of-fit
□ Monotonicity measures

Parameters

☑ Estimates
☑ Likelihood ratio tests
□ Asymptotic correlations
□ Asymptotic covariances

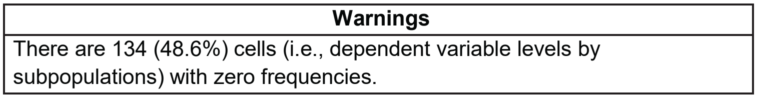
Confidence Interval (%): 95

Define Subpopulations

● Covariate patterns defined by factors and covariates
○ Covariate patterns defined by variable list below

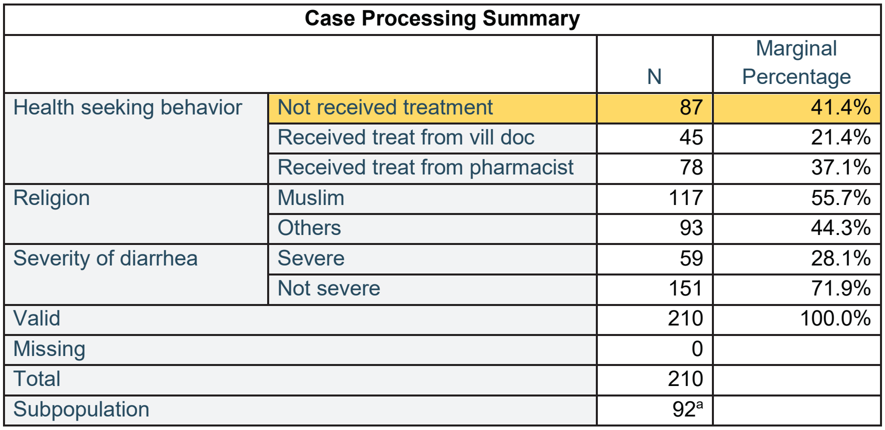
15.2 Outputs

Table 15.1 Warning message



Warnings
There are 134 (48.6%) cells (i.e., dependent variable levels by subpopulations) with zero frequencies.

Table 15.2 Case Processing Summary



Case Processing Summary			
		N	Marginal Percentage
Health seeking behavior	Not received treatment	87	41.4%
	Received treat from vill doc	45	21.4%
	Received treat from pharmacist	78	37.1%
Religion	Muslim	117	55.7%
	Others	93	44.3%
Severity of diarrhea	Severe	59	28.1%
	Not severe	151	71.9%
Valid		210	100.0%
Missing		0	
Total		210	
Subpopulation		92[a]	

a. The dependent variable has only one value observed in 50 (54.3%) subpopulations.

Table 15.3 Model Fitting Information

Model Fitting Information				
	Model Fitting Criteria	Likelihood Ratio Tests		
Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	323.351			
Final	226.663	96.687	6	.000

Table 15.4 Classification table

Classification				
	Predicted			
Observed	Not received treatment	Received treat from vill doc	Received treat from pharmacist	Percent Correct
Not received treatment	48	9	30	55.2%
Received treat from vill doc	3	34	8	75.6%
Received treat from pharmacist	27	6	45	57.7%
Overall Percentage	37.1%	23.3%	39.5%	60.5%

Table 15.5 Goodness-of-Fit

Goodness-of-Fit			
	Chi-Square	df	Sig.
Pearson	153.191	176	.892
Deviance	144.880	176	.958

Table 15.6 Pseudo R-Square

Pseudo R-Square	
Cox and Snell	.369
Nagelkerke	.419
McFadden	.217

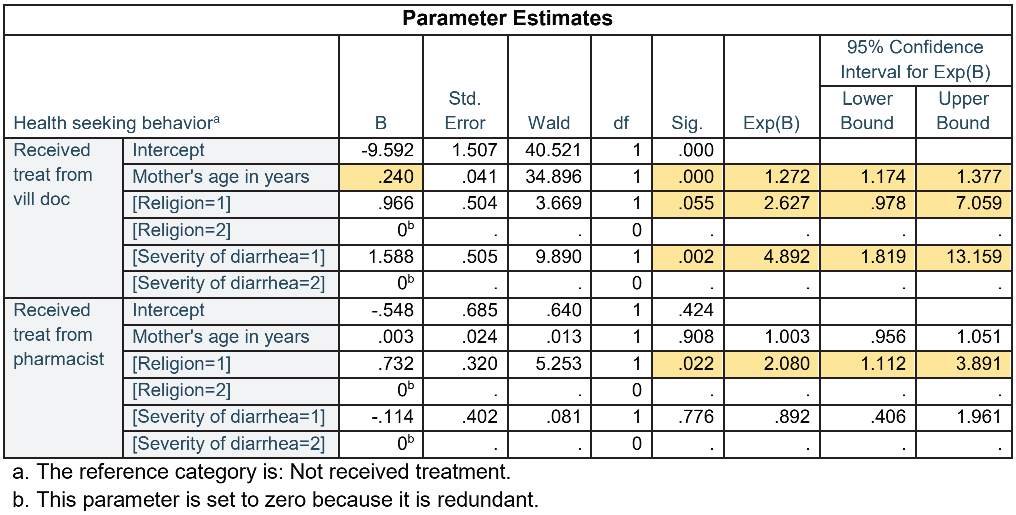
Table 15.7 Likelihood Ratio Tests

Likelihood Ratio Tests				
	Model Fitting Criteria	Likelihood Ratio Tests		
Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	226.663[a]	.000	0	.
Mother's age in years	289.706	63.043	2	.000
Religion	233.729	7.066	2	.029
Severity of diarrhea	239.900	13.236	2	.001

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Table 15.8 Parameter Estimates



Parameter Estimates

Health seeking behavior[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower Bound	95% Confidence Interval for Exp(B) Upper Bound
Received treat from vill doc	Intercept	-9.592	1.507	40.521	1	.000			
	Mother's age in years	.240	.041	34.896	1	.000	1.272	1.174	1.377
	[Religion=1]	.966	.504	3.669	1	.055	2.627	.978	7.059
	[Religion=2]	0[b]	.	.	0	.	.	.	.
	[Severity of diarrhea=1]	1.588	.505	9.890	1	.002	4.892	1.819	13.159
	[Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.
Received treat from pharmacist	Intercept	-.548	.685	.640	1	.424			
	Mother's age in years	.003	.024	.013	1	.908	1.003	.956	1.051
	[Religion=1]	.732	.320	5.253	1	.022	2.080	1.112	3.891
	[Religion=2]	0[b]	.	.	0	.	.	.	.
	[Severity of diarrhea=1]	-.114	.402	.081	1	.776	.892	.406	1.961
	[Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Not received treatment.

b. This parameter is set to zero because it is redundant.

15.3 Interpretation

In the outputs, we have the warning message at the beginning (Table 15.1). This warning message usually appears if there is any quantitative variable included in the model. The warning message indicates that there are some (n=134, in our example) cells with zero frequencies. If you include only the categorical variables in the model, you may not get this warning message (provided there is no cell with zero frequency). However, this warning message does not have any effect on the analysis.

Table 15.2 shows that there are three levels/categories in the dependent variable (health seeking behavior), including the number of subjects in each category with their percentages (marginal percentage). The table also shows that there are two categorical variables (religion and severity of diarrhea) in the model, including the number of subjects in each category. For example, the variable "religion" shows that a total of 117 (55.7%) individuals were Muslim and the rest of the 93 individuals (44.3%) were from other religions.

Table 15.3 (model fitting information) provides information on whether addition of independent variables in the model has improved the ability to predict the outcome compared to the null model. The null model considers the modal (most frequent) class of the dependent variable (i.e., "not received treatment" in our example) as the model's prediction accuracy (41.4%; Table 15.2). We expect that the p-value of the "Final (last row in the table)" is significant (i.e., p-value is <0.05). Here the p-value is 0.000, which indicates that the addition of the variables "religion" and "severity of diarrhea" have improved the ability to predict the outcome.

The improvement of the model can also be checked if we compare the "Percent processing summary table (Table 15.2) at different levels of the dependent variable. Table 15.4 shows that the "Percent Correct" has increased when compared with the "Marginal Percentage" (Table 15.2) at each level of the dependent variable, such as did not receive treatment (55.2% vs 41.4%), received treatment from village doctor (75.5% vs 21.4%) and received treatment from pharmacist (57.7% vs 37.1%). This indicates that our model compared to the null model gives better accuracies for all the groups (i.e., not received treatment, received treatment from the village doctor and received treatment from the pharmacist).

Table 15.5 indicates whether the model is good for prediction or not. If the model is good for prediction, the p-values (Sig.) will be >0.05. Since the p-values (Sig.) of Goodness-of-fit tests (Pearson and Deviance) are >0.05, we can conclude that this model adequately fits the data. If there are many sub-populations (cells) with zero frequencies, the p-values of the test may be <0.05. In that case, the Goodness-of-fit test is not important and you can just ignore it. *Note that all the information provided in Tables 15.3-15.5 are needed if our interest is in prediction.*

The "Pseudo R-square" values, as shown in Table 15.6, indicate how much variation in the dependent variable can be explained by the independent variables in the model. The results show that 21.7% to 41.9% variation in the dependent variable can be explained by the independent variables (age, religion and severity of diarrhea) in the model.

The Likelihood ratio test (Table 15.7) indicates the contribution of each variable to the model. Table 15.7 shows that all the variables (age, religion and severity of diarrhea) have significant contribution to the model since all the p-values are <0.05.

Table 15.8 (Parameter estimates) is the main table for interpretation of the results. The first half of the table has the results for "received treatment from village doctor" compared to "did not seek treatment". The results indicate that Muslims (compared to other religions) are more likely to receive treatment from the village doctors after controlling for mothers' age and severity of diarrhea, but the association is not statistically significant (adjusted OR: 2.62; 95% CI: 0.97-7.05; p=0.055). However, severity of diarrhea (i.e., if the baby has severe diarrhea) is significantly associated with seeking treatment from the village doctors after adjusting for mothers' age and religion (adjusted OR: 4.89; 95% CI: 1.81-13.15; p=0.002).

For the quantitative variables, positive coefficients (indicated by B in Table 15.8) indicate an increased likelihood of the response category (received treatment from village doctors) compared to the reference category (did not seek treatment). The results show that with the increase in mothers' age, it is significantly more likely to receive treatment from the village doctors after adjusting for religion and severity of diarrhea (adjusted OR: 1.27; 95% CI: 1.17-1.37; p=0.000) [in other words, there is a 27% increase in odds with each year increase in mothers' age].

The second half of the table shows the results for “received treatment from the pharmacists” compared to “did not seek treatment”. The interpretations are similar as mentioned above. The results show that only religion (being Muslim) is significantly associated with “received treatment from the pharmacists” after controlling for maternal age and severity of diarrhea (adjusted OR: 2.08; 95% CI: 1.11-3.89; p=0.022).

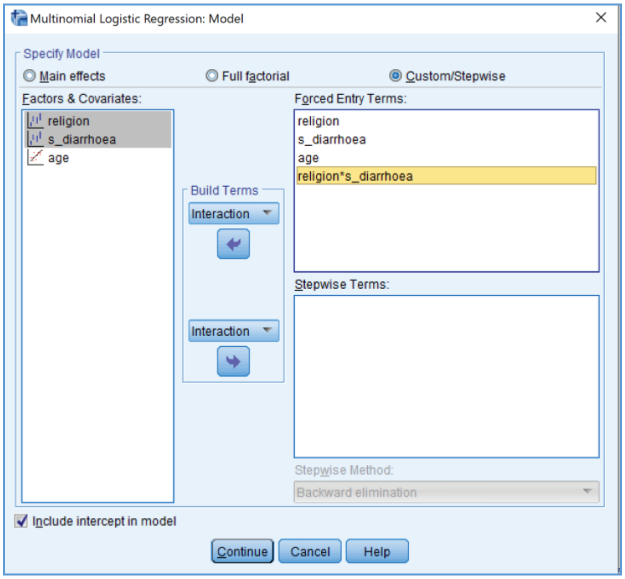
15.4 Incorporating interaction terms in the multinominal logistic regression model

If you want to check interactions along with the main effects, you need to use the “*Customs/Stepwise*” method under the “Model” option. In this example, we shall see how to include the interaction term between religion and severity of diarrhea in the model. To perform the analysis, use the following commands.

> Analyze > Regression > Multinominal Logistic > Select the variable “behavior” and push it into the “Dependent” box > Click on “Reference Category” > Select “First category” under “Reference Category” > Continue > Select the variables “religion” and “s_diarrhoea” and push them into the “Factor(s)” box > Select “age” and push it into the “Covariate(s)” box > Click on “Model” > Select “Customs/Stepwise” > Select “Main Effects” under “Build Terms” clicking the down arrow > Select the variables “religion” and “s_diarrhoea” in “Factors and Covariates” box and push them into the “Forced Entry Terms” box > Now select “Interaction” under “Build Terms” clicking the down arrow > Select “religion” and “s_diarrhoea” together in “Factors and Covariates” box pressing the “Ctrl” button of the key board > Push them into the “Forced Entry Terms” box (you will see “religion*s_diarrhoea” in the box) [Fig 15.3] > Continue > Click on “Statistics” > Select “Pseudo R-square, Step summary, Model fitting information, Classification table, and Goodness-of-fit” under “Model” > Select “Estimates and Likelihood ratio tests” under “Parameters” > Select “Covariate patterns defined by factors and covariates” under “Define Subpopulations” > Continue > OK

The above commands will provide the following table (only the main table for interpretation of the results is shown) of parameter estimates (Table 15.9). The table shows that there is no interaction between religion and severity of diarrhoea for both the outcome categories (p-values are 0.847 and 0.551, respectively).

Figure 15.3



Multinomial Logistic Regression: Model

Specify Model

○ Main effects ○ Full factorial ● Custom/Stepwise

Factors & Covariates:

- religion
- s_diarrhoea
- age

Build Terms

Interaction

Interaction

Forced Entry Terms:

- religion
- s_diarrhoea
- age
- religion*s_diarrhoea

Stepwise Terms:

Stepwise Method: Backward elimination

☑ Include intercept in model

Continue Cancel Help

Table 15.9 Parameter Estimates

Health seeking behavior[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower Bound	95% Confidence Interval for Exp(B) Upper Bound
Received treat from vill doc	Intercept	-1.792	.408	19.262	1	.000			
	[Religion=1]	.857	.542	2.507	1	.113	2.357	.815	6.813
	[Religion=2]	0[b]	.	.	0	.	.	.	.
	[Severity of diarrhea=1]	1.792	.673	7.097	1	.008	6.000	1.606	22.421
	[Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.
	[Religion=1] * [Severity of diarrhea=1]	-.164	.854	.037	1	.847	.848	.159	4.523
	[Religion=1] * [Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.
	[Religion=2] * [Severity of diarrhea=1]	0[b]	.	.	0	.	.	.	.
	[Religion=2] * [Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.
Received treat from pharmacist	Intercept	-.519	.253	4.218	1	.040			
	[Religion=1]	.824	.355	5.398	1	.020	2.280	1.138	4.570
	[Religion=2]	0[b]	.	.	0	.	.	.	.
	[Severity of diarrhea=1]	.182	.638	.082	1	.775	1.200	.344	4.188
	[Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.
	[Religion=1] * [Severity of diarrhea=1]	-.488	.818	.356	1	.551	.614	.124	3.050
	[Religion=1] * [Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.
	[Religion=2] * [Severity of diarrhea=1]	0[b]	.	.	0	.	.	.	.
	[Religion=2] * [Severity of diarrhea=2]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Not received treatment.

b. This parameter is set to zero because it is redundant.

16

Survival Analysis

In many situations, researchers are interested to know the progress of a patient (with a disease) from a specific point in time (e.g., from the point of diagnosis or from the point of initiation of treatment) until the occurrence of a certain outcome, such as death or recurrence of any event (e.g., recurrence of cancer). Prognosis of a condition is usually assessed by estimating: a) the *median survival time*, and b) the *cumulative probability of survival* after a certain time interval (e.g., 5-year, 3-year).

For example, a researcher may be interested to know what is the median survival time of colonic cancer if the patient is not treated (or treated), and what is the estimated probability that a patient may survive for more than 5 years (5-year cumulative survival probability), if the patient is treated (or not treated). The methods employed to answer these questions in a follow-up study are known as survival analysis (or life table analysis) methods.

Survival analysis is done in the follow-up studies. To do the survival analysis, we need to have data (information) from each of the patients, at least on:

- *Time*: Length of time the patient was observed in the study (called survival time);
- *Outcome*: Whether the patient developed the outcome of interest (event) during the study period, or the patient was either lost to follow-up or remained alive at the end of the study (censored); and
- *Treatment group*: Which treatment (e.g., treatment A or B) did the patient receive in the study (optional)?

The survival time is of two types – a) Censored time; and b) Event time. The *censored* time is the amount of time contributed by:

a) The patients who did not develop the outcome and remained in the study up to the end of the study period; or
b) Patients who were lost to follow-up due to any reason, such as migration and withdraw; or

c) Patients who developed the outcome (e.g., died due to accident) due to other reasons than the disease of interest.

On the other hand, the *event* time is the amount of time contributed by the patients who developed the outcome of interest during the study period.

If we have the above information, it is possible to estimate the median survival times and cumulative survival probabilities for two or more treatment groups for comparison. Such a comparison allows us to answer the question "which treatment delays the time of occurrence of the event". The method commonly used to analyse the survival-time data is the *Kaplan-Meier* method, and SPSS can be used for the analysis of such data. Use the data file <**Data_survival_4.sav**> for practice.

16.1 Survival analysis: Kaplan-Meier method

Assume that a researcher has conducted a follow-up study (clinical trial) on patients with heart failure to determine the effectiveness of a new drug (n=22) compared to a placebo (n=22). The outcome of interest in this study is death (event). The objective is to assess whether the "new treatment" delays the time to death (event) compared to placebo among the patients with heart failure. Following variables are included in the data file.

- *Time*: It is the amount of time each patient has spent in the study in days;
- *Treatment*: Which treatment did the patient receive (0= placebo; 1= new treatme--nt);
- *Outcome (event)*: Whether the patient developed the event, i.e., died or not (0 = censored; 1= died)

Assumptions

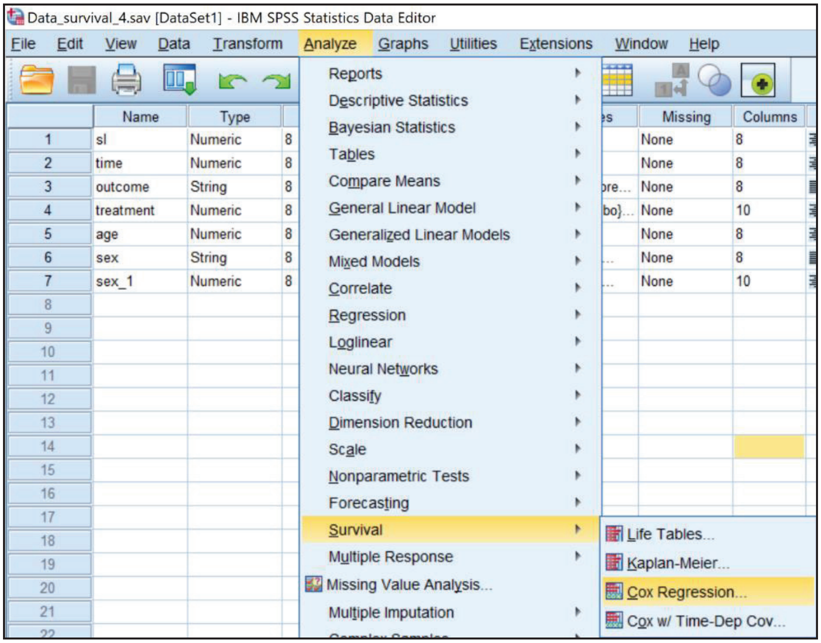
- The probability of the outcome is similar among the censored and under-observa--tion individuals;
- There is no secular trend over the calendar period;
- The risk is uniform during the interval;
- Loss to follow-up is uniform over the interval.

16.1.1 Commands

Analyze > Survival > Kaplan Meier > Push the variable "time" to "Time" box > Push the variable "outcome" in the "Status" box > Click "Define event" > Select "Single value" and type "1" (here 1 is the event) in the box > Continue > Push "treatment" in the "Factor" box > Click Options… > Select "Survival table(s), and "Mean and Med-

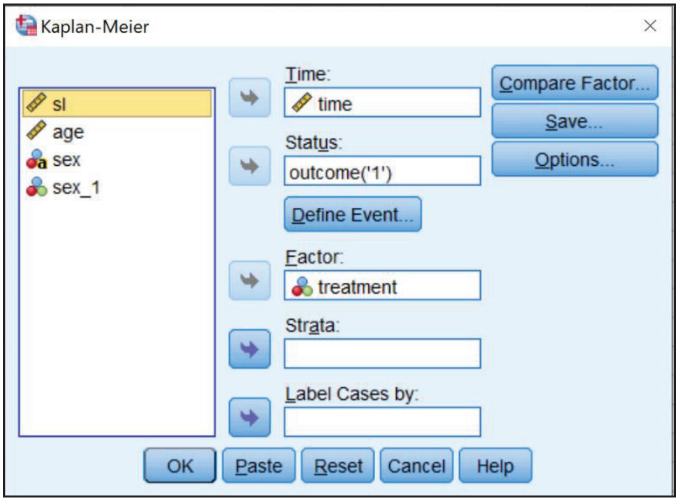
-ian survival” under statistics > Select “Survival” under “Plots” > Continue > Click “Compare Factor…” > Select “Log rank” under “Test statistics” > Continue > OK (Figs 17.1 and 17.2)

Figure 16.1



	Name	Type		es	Missing	Columns
1	sl	Numeric	8		None	8
2	time	Numeric	8		None	8
3	outcome	String	8	pre...	None	8
4	treatment	Numeric	8	bo}...	None	10
5	age	Numeric	8		None	8
6	sex	String	8	...	None	8
7	sex_1	Numeric	8	...	None	10

Figure 16.2



Time: time

Status: outcome('1')

Factor: treatment

Strata:

Label Cases by:

16.1.2 Outputs

The SPSS will give the following outputs.

Table 16.1 Case Processing Summary

Treatment group	Total N	N of Events	Censored	
			N	Percent
Placebo	22	16	6	27.3%
New treatment	22	11	11	50.0%
Overall	44	27	17	38.6%

Table 16.2 Means and Medians for Survival Time

Treatment group	Mean[a]				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Placebo	72.545	14.839	43.462	101.629	40.000	12.899	14.719	65.281
New treatment	125.264	13.402	98.996	151.532	146.000	28.786	89.580	202.420
Overall	98.925	10.812	77.733	120.117	89.000	21.232	47.385	130.615

a. Estimation is limited to the largest survival time if it is censored.

Table 16.3 Overall Comparisons (Log Rank test results)

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	4.660	1	.031

Test of equality of survival distributions for the different levels of Treatment group.

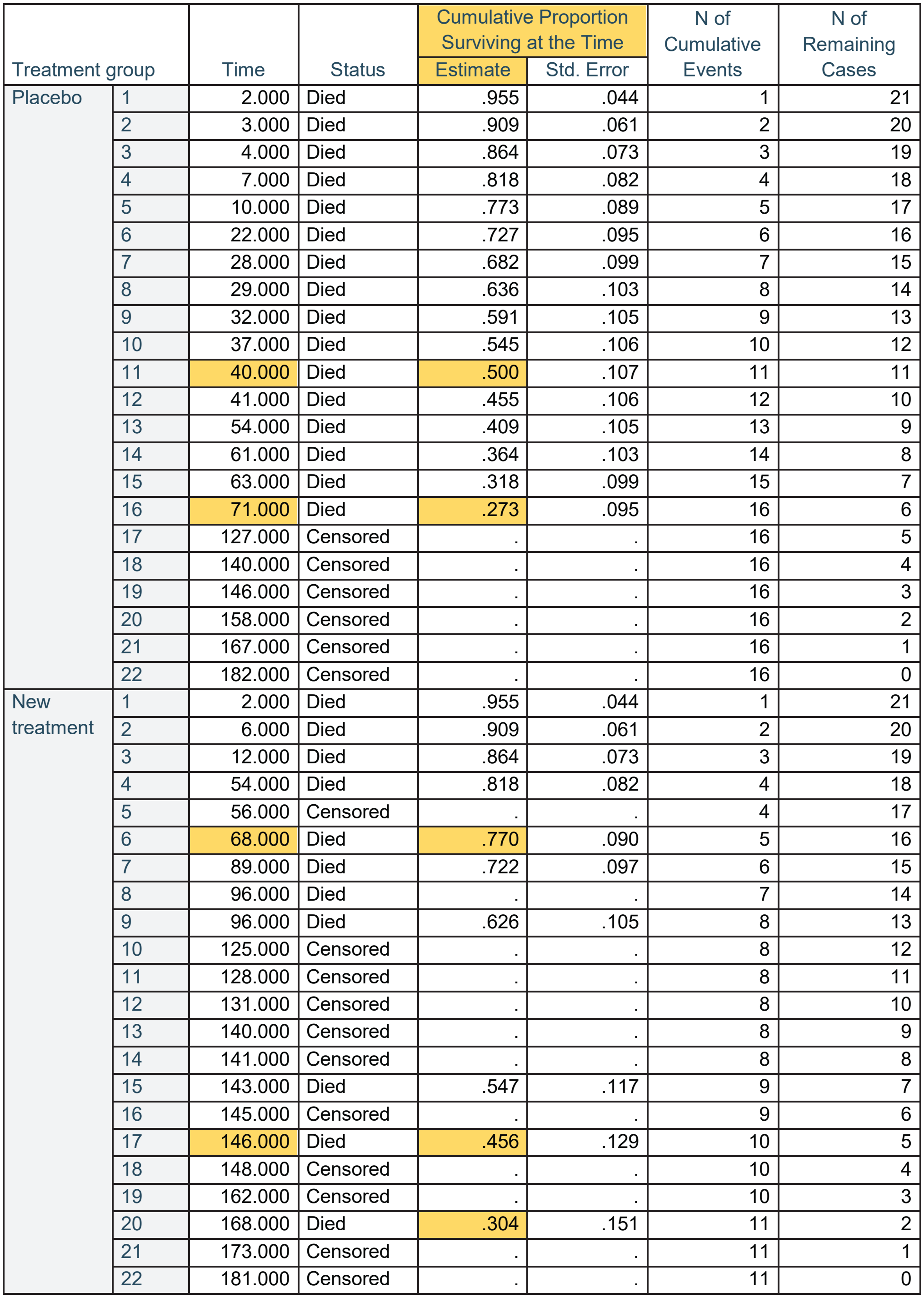
Table 16.4 Log Rank, Breslow, and Tarone-Ware test results

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	4.660	1	.031
Breslow (Generalized Wilcoxon)	6.543	1	.011
Tarone-Ware	6.066	1	.014

Test of equality of survival distributions for the different levels of Treatment group.

Table 16.5 Survival Table



Treatment group		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
Placebo	1	2.000	Died	.955	.044	1	21
	2	3.000	Died	.909	.061	2	20
	3	4.000	Died	.864	.073	3	19
	4	7.000	Died	.818	.082	4	18
	5	10.000	Died	.773	.089	5	17
	6	22.000	Died	.727	.095	6	16
	7	28.000	Died	.682	.099	7	15
	8	29.000	Died	.636	.103	8	14
	9	32.000	Died	.591	.105	9	13
	10	37.000	Died	.545	.106	10	12
	11	40.000	Died	.500	.107	11	11
	12	41.000	Died	.455	.106	12	10
	13	54.000	Died	.409	.105	13	9
	14	61.000	Died	.364	.103	14	8
	15	63.000	Died	.318	.099	15	7
	16	71.000	Died	.273	.095	16	6
	17	127.000	Censored	.	.	16	5
	18	140.000	Censored	.	.	16	4
	19	146.000	Censored	.	.	16	3
	20	158.000	Censored	.	.	16	2
	21	167.000	Censored	.	.	16	1
	22	182.000	Censored	.	.	16	0
New treatment	1	2.000	Died	.955	.044	1	21
	2	6.000	Died	.909	.061	2	20
	3	12.000	Died	.864	.073	3	19
	4	54.000	Died	.818	.082	4	18
	5	56.000	Censored	.	.	4	17
	6	68.000	Died	.770	.090	5	16
	7	89.000	Died	.722	.097	6	15
	8	96.000	Died	.	.	7	14
	9	96.000	Died	.626	.105	8	13
	10	125.000	Censored	.	.	8	12
	11	128.000	Censored	.	.	8	11
	12	131.000	Censored	.	.	8	10
	13	140.000	Censored	.	.	8	9
	14	141.000	Censored	.	.	8	8
	15	143.000	Died	.547	.117	9	7
	16	145.000	Censored	.	.	9	6
	17	146.000	Died	.456	.129	10	5
	18	148.000	Censored	.	.	10	4
	19	162.000	Censored	.	.	10	3
	20	168.000	Died	.304	.151	11	2
	21	173.000	Censored	.	.	11	1
	22	181.000	Censored	.	.	11	0

16.1.3 Interpretation

Table 16.1 is the summary table indicating the number of study subjects in each group (22 in the placebo and 22 in the new treatment group) and the number of events (number of deaths) occurred in each group including the number censored. The table shows that in the treatment group, 11 patients died and 11 were censored, while in the placebo group, 16 died and 6 were censored.

Table 16.2 shows the mean and median survival times for both the placebo and new treatment groups. We do not consider the mean survival time for reporting. We consider the *median survival time*. The median survival time is the time when the cumulative survival probability is 50%. The table indicates that the median survival time, if the patient is in the placebo group, is 40 days (95% CI: 14.71 to 65.28), while it is 146 days (95% CI: 89.58 to 202.42), if the patient is in the new treatment group. This means that the new treatment increases the survival time, i.e., the new treatment is associated with longer time-to-event (and the placebo is associated with shorter time-to-event). Therefore, we conclude that the person lives longer if s/he receives the new treatment compared to the placebo.

Table 16.5 shows the survival probability (Cumulative Proportion Surviving at the Time) at different points in time in the placebo and treatment group. In this table, we can see that the cumulative survival probability at the end of 71 days (time column), in the placebo group, is 0.273 (27.3%). Since there is no death after that, the cumulative survival probability at the end of 182 days will be the same (0.273).

On the other hand, the cumulative survival probability is 0.304 (30.4%) at the end of 168 days, if the patient is in the new treatment group. As there is no death after that, the cumulative survival probability at the end of 181 days will be the same (0.304). In the new treatment group, the cumulative survival probability at the end of 71 days (68 days in the table) is about 0.770 (77.0%), which is much higher than the placebo group (0.273). This indicates that the probability of survival at the end of 71 days is higher among the patients who received the new treatment compared to placebo. This also indicates the benefit of the new treatment (i.e., the new treatment is better than the placebo).

However, if we consider the cumulative survival probability of patients in both these groups at the end of 180 days, the outcome is not that different – 0.273 in the placebo group and 0.304 in the treatment group. This information suggests that though the difference is small, the survival probability is still higher if the person is on the new treatment than on the placebo.

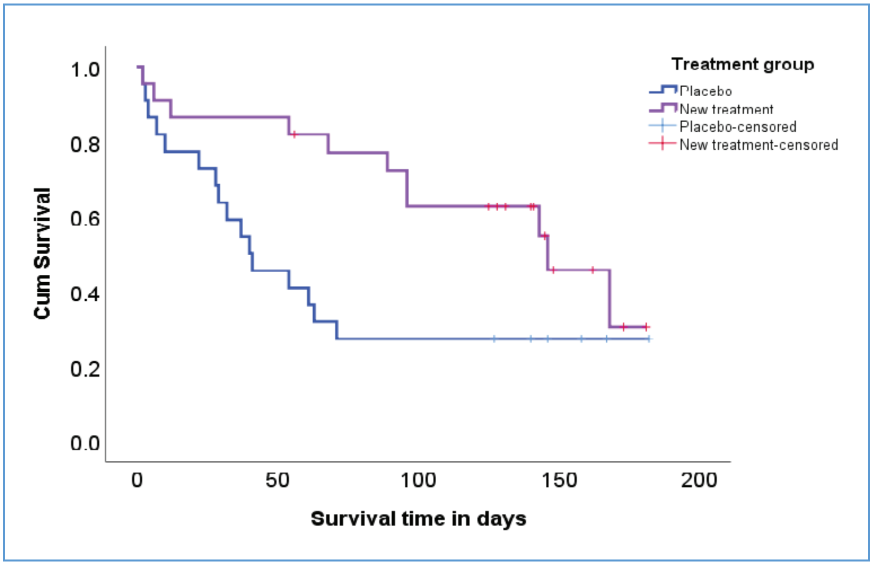
We can also estimate the median survival time (it is the time when the cumulative survival probability is 50%) in both these groups from this table. The median survival time for the placebo group is 40 days and that of the treatment group is 146 days (Table 16.5). Now, the question is whether the survival experiences of both these groups in the population are different or not? To answer this question, we have to use a statistical test (Log Rank test) as given in Table 16.3.

Table 16.3 shows the *Log Rank* test results. For an objective comparison of the survival experience of two groups, it is desirable to use some statistical methods that will tell us whether the difference of the survival experiences in the population is statistically significant or not. Here, the null hypothesis is “there is no difference in the survival experience of the two groups (new treatment and placebo) in the population”. Such a null hypothesis is tested by the *Log Rank* test. The Log Rank test results show that the p-value is 0.031, which is <0.05. This means that the survival experience of both these groups in the population is not same. In other words, it indicates that the survival probability is better if the patient receives the new treatment (i.e., the new treatment is more effective/better than the placebo in improving the patients’ survival, since the median survival time is higher in the new treatment group).

Note that there are alternative procedures for testing the null hypothesis that the two survival curves are identical. They are the *Breslow test, Tarone-Ware test and Peto test* (Table 16.4). The Log Rank test ranks all the deaths equally, while the other tests give more weight to early deaths. The options are available in SPSS under the “Compare Factor” tab.

Survival curve: The cumulative survival probability is usually portrayed visually by a graph called survival curve (Fig 16.3). The “steps” in the graph represent the time when events (deaths or any other event of interest) occurred. The graph allows us to represent visually the median survival time and the cumulative survival probability for any specific time period (e.g., 30-day; 6-month; 1-year, 3-year, 5-year, 10-year cumulative survival probability). In general, the line above indicates the better survival probability. We can see that the line for the new treatment is above the line for the placebo.

Figure 16.3 Survival curve



17

Cox Regression

The Cox regression is also called *Proportional Hazards Analysis*. In the previous chapter (chapter 16), we have discussed the survival analysis using the Kaplan Meier method. Like other regression methods (e.g., multiple linear regression and logistic regression), Cox regression is a multivariable analysis technique where the dependent measure is a mixture of time-to-event and censored-time observations. The Cox regression is commonly done in follow-up studies (e.g., randomized trials) to assess the prognosis. The Cox regression with constant time can also be used for the analysis of cross-sectional data to estimate the prevalence ratio. Use the data file <**Data_survival_4.sav**> for practice.

17.1 Cox regression or proportional hazards analysis

Returning to our previous example (chapter 16) where we have analyzed the data to assess the effectiveness of a new treatment compared to the placebo. Our objective was to determine whether the new treatment delays the time-to-death compared to the placebo among patients with heart failure. We found that the new treatment significantly delayed the time-to-death compared to placebo, as indicated by the higher median survival time and a significant Log Rank test. However, the effectiveness of the new treatment might be influenced (confounded) by other factors, such as age, hypertension, diabetes or other characteristics. All these variables, therefore, need to be controlled during analysis for assessing the effectiveness of the new treatment. The Cox regression is a statistical method that is used to control the confounding factors (categorical, continuous or discrete covariates) that may influence the effectiveness of the new treatment.

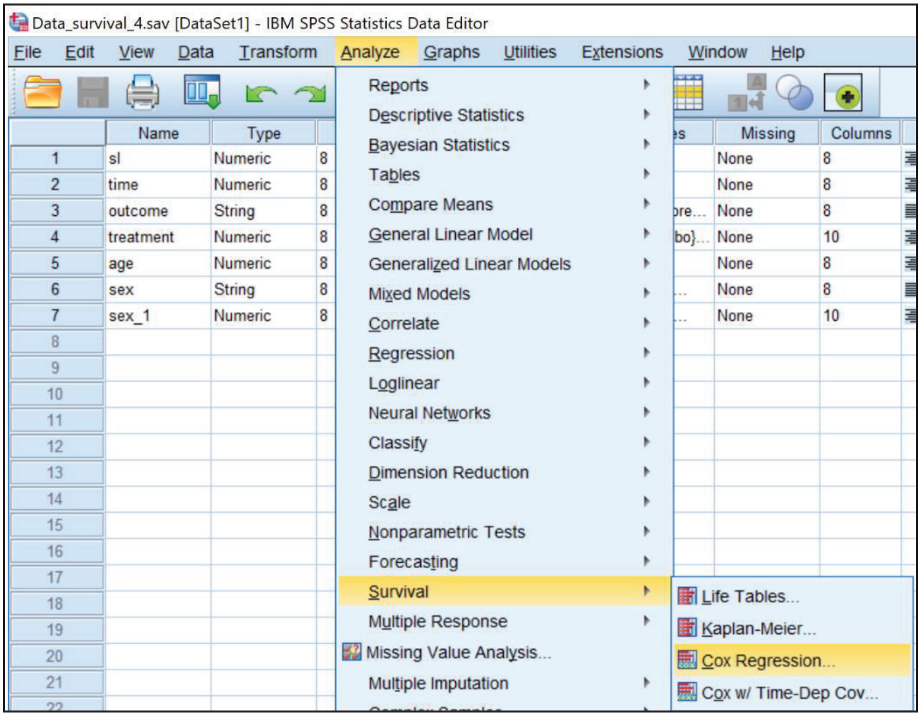
The Cox regression gives us the *Hazard Ratio*, which is analogous to Relative Risk (RR). Hazard Ratio (also called Relative Hazard) is the ratio of hazards if the persons are exposed compared to the persons not exposed. In Cox regression, the dependent variable is the Log of hazard.

17.1.1 Commands

Let us use the previous example and data for the Cox regression analysis along with the variables "sex" and "age" for adjustment. Note that the variable "treatment" has two categories – placebo (coded as "0") and new treatment (coded as "1").

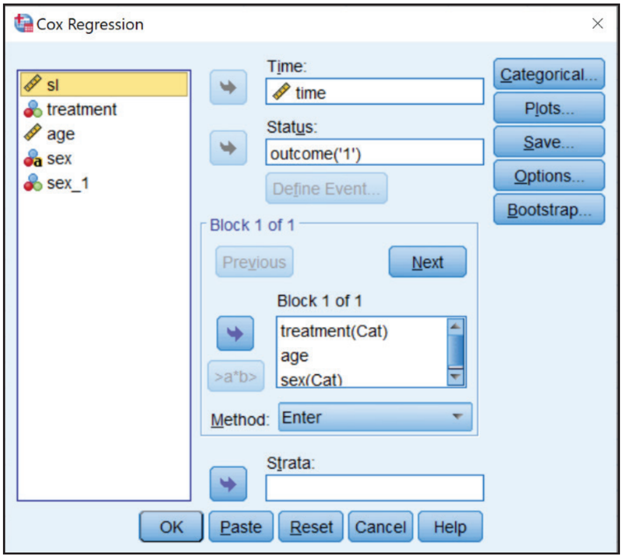
> Analyze > Survival > Cox Regression > Push "time" into the "Time" box > Push "outcome" into the "Status" box > Click on "Define event" > In "Single value" box write 1 (since 1 is the code no. of the event) > Continue > Push "treatment, age and sex" into the "Covariate" box > Click on "Categorical" > Push "treatment and sex" into the "Categorical covariates" box > Select "treatment" in the "Categorical covariates" box > Select "Last" from "Reference category" (usually the default) under "Change contrast" (we are doing this to make "new treatment" as the comparison group) > Change > Continue > Click on "Options" > Select "CI for Exp(B) and Correlation of estimates" > Continue > Click on "Plots" > Select "Survival and Log minus Log" > Select the variable "treatment" from the "Covariate Values Plotted at" and push into the "Separate Line for" box > Continue > OK (Figs 17.1 to 17.4)

Figure 17.1



	Name	Type		es	Missing	Columns
1	sl	Numeric	8		None	8
2	time	Numeric	8		None	8
3	outcome	String	8	pre...	None	8
4	treatment	Numeric	8	bo}...	None	10
5	age	Numeric	8		None	8
6	sex	String	8	...	None	8
7	sex_1	Numeric	8	...	None	10
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						

Figure 17.2


Cox Regression

sl
treatment
age
sex
sex_1

Time: time

Status: outcome('1')

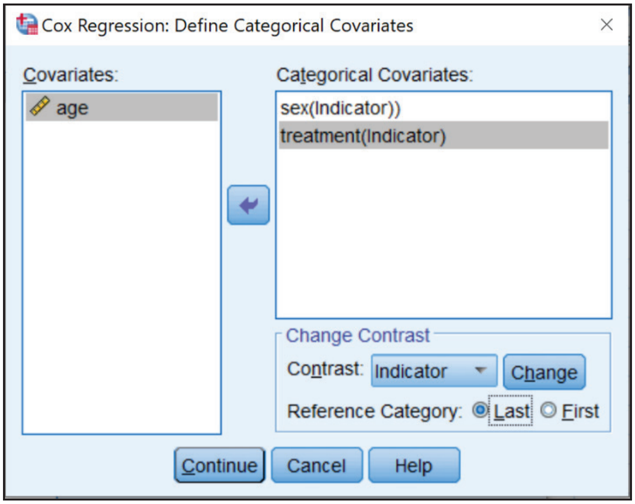
Block 1 of 1

Block 1 of 1
treatment(Cat)
age
sex(Cat)

Method: Enter

Strata:

Figure 17.3


Cox Regression: Define Categorical Covariates

Covariates:
age

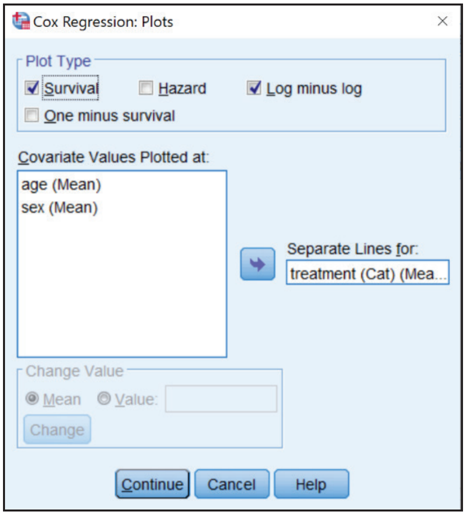
Categorical Covariates:
sex(Indicator))
treatment(Indicator)

Change Contrast

Contrast: Indicator

Reference Category: ● Last ○ First

Figure 17.4



Cox Regression: Plots

Plot Type

☑ Survival □ Hazard ☑ Log minus log

□ One minus survival

Covariate Values Plotted at:

age (Mean)

sex (Mean)

Separate Lines for: treatment (Cat) (Mea...

Change Value

● Mean ○ Value:

17.1.2 Outputs

Only the relevant tables are provided below.

Table 17.1 Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	27	61.4%
	Censored	17	38.6%
	Total	44	100.0%
Cases dropped	Cases with missing values	0	0.0%
	Cases with negative time	0	0.0%
	Censored cases before the earliest event in a stratum	0	0.0%
	Total	0	0.0%
Total		44	100.0%

a. Dependent Variable: Survival time in days

Table 17.2 Categorical Variable Codings[a,d]

		Frequency	(1)[c]
Treatment group[b]	0=Placebo	22	1
	1=New treatment	22	0
Sex[b]	0=male	21	1
	1=female	23	0

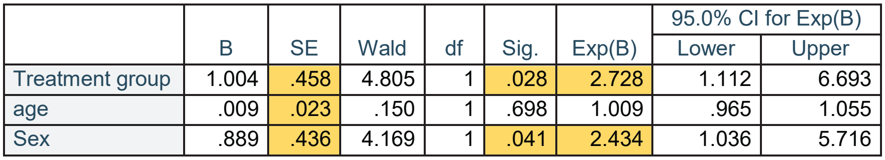
a. Category variable: Treatment group (treatment)

b. Indicator Parameter Coding

c. The (0,1) variable has been recoded, so its coefficients will not be the same as for indicator (0,1) coding.

d. Category variable: Sex (sex)

Table 17.3 Variables in the Equation



	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Treatment group	1.004	.458	4.805	1	.028	2.728	1.112	6.693
age	.009	.023	.150	1	.698	1.009	.965	1.055
Sex	.889	.436	4.169	1	.041	2.434	1.036	5.716

Figure 17.5 Survival function for patterns 1 – 2

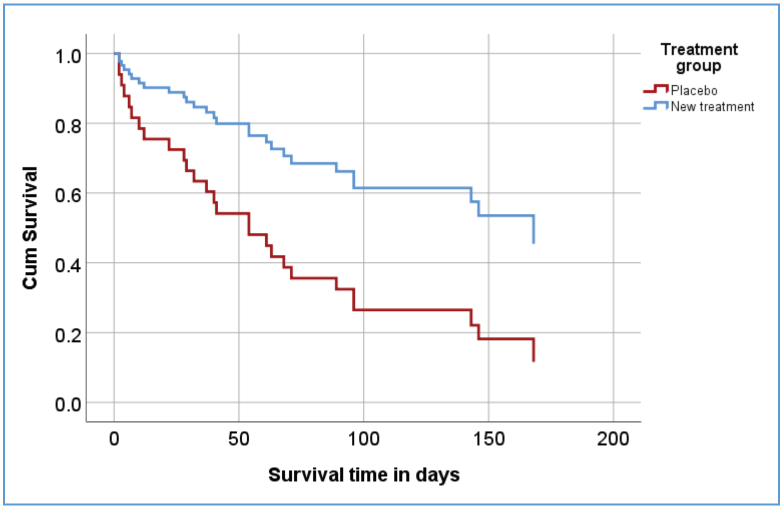
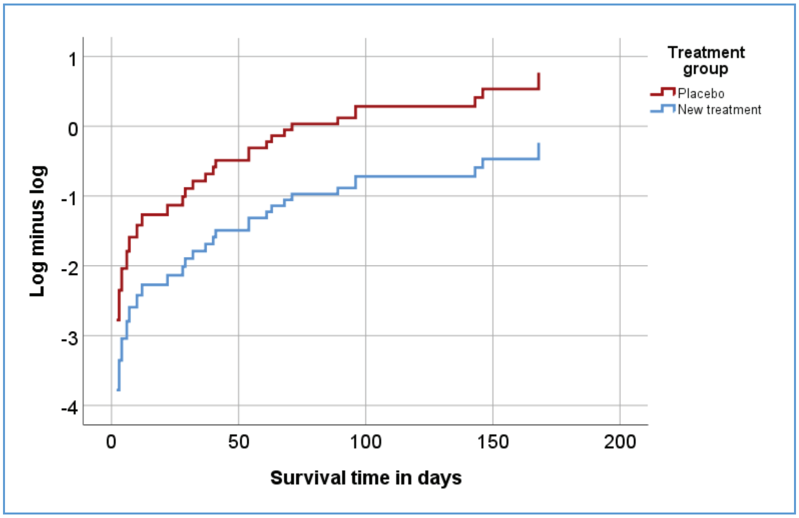


Figure 17.6 Log minus log (LML) function for patterns 1 – 2



17.1.3 Interpretation

Table 17.1 shows the number of cases that are analyzed. Table 17.2 is *very important* for interpretation. This table indicates which category of the categorical variables is the comparison group. Look at the last column [(1)^b] of the table. The value "0" in this column indicates the comparison group. In the table (Table 17.2), the "New treatment" is indicated as "0" in the last column. Therefore, in the analysis, "new treatment" is the comparison group (though "new treatment" is actually coded as "1"). Similarly, females are the comparison group in this analysis since the value of being "female" is "0" in the last column. We shall, therefore, get the Hazard Ratio for the "placebo" group compared to the "new treatment" group and for the "males" compared to the "females", as shown in Table 17.3 (Variables in the Equation).

Our main interest is in Table 17.3 (Variables in the Equation). The table indicates the Hazard Ratio [Exp(B)], p-value (Sig.) and 95% confidence interval (CI) for the *Hazard Ratio* [95% CI for Exp(B)]. The Hazard Ratio for the variable "treatment" is 2.72 (95% CI: 1.11 to 6.69) and the p-value is 0.028 (note that the "new treatment" is the comparison group. Therefore, the results provided by SPSS are for the "placebo" group compared to the "new treatment" group). This indicates that compared to the "new treatment", patients in the "placebo" group are 2.72 times more likely to *have shorter time to event* after controlling for "age" and "sex", which is statistically significant (p=0.028) at 95% confidence level. On the other hand, males are more likely (2.43 times) to have shorter time to event compared to the females after controlling for the variables "treatment" and "age" (p=0.041). Age, independently, does not have any significant effect on the survival time, since the p-value is 0.698.

Figure 17.5 shows the survival plot/curve of the heart failure patients by treatment group. The upper line is for the "new treatment" group and the lower one is for the "placebo" group. The figure shows the outcome difference between the new treatment and placebo. The group represented by the upper line has the better survival probability.

However, before we conclude the results, we have to check if: a) there is a multicollinearity among the independent variables; and b) relative hazards over the time are proportional (also called the proportionality assumption of the proportional hazards analysis). Look at the SE of the variables in the model (Table 17.3). There is no value which is very small (<0.001) or very large (>5.0) (refer to the logistic regression analysis, section 14.1.5), indicating that there is no problem of multicollinearity in the model.

For the second assumption, we need to check the *log-minus-log* survival plot (Fig 17.6). If there is a constant vertical difference between the two curves (i.e., the curves are parallel to each other), it means that the relative hazards over the time are proportional. If the curves cross each other, or are much closer together at some points in time and much further apart at other points in time, then the assumption is violated. In our example two lines are more or less parallel, indicating that the assumption is not violated. When the proportional hazards assumption is violated, it is recommended to use the Cox regression with time dependent covariate to analyze the data.

Finally, like multiple linear regression and logistic regression analyses, we can use the automatic selection method of independent variables for modeling. To do this, select "Backward LR" for "Method" from the template 17.2 (Fig 17.2).

17.2 Proportional hazards analysis with constant time

It is a common practice to analyse data of a cross-sectional study using logistic regression when the outcome variable is dichotomous. The logistic regression model, when used for cross-sectional data, provides the adjusted OR (prevalence OR), not the prevalence ratio (PR). In cross-sectional studies, when the prevalence is more than 10%, OR overestimates the PR if PR>1 (i.e., if PR>1, the POR is always >PR). Controlling for the confounding factors is not also equivalent for these two measures. Thus, PR should be used in preference to the OR. It is, therefore, recommended to use the proportional hazards model (Cox regression) with constant time, which provides statistical methods directed specifically at the PR. The other option is to use the generalized linear model (log-binomial regression). Here, we have discussed how to analyse the cross-sectional data and interpret the results using proportional hazards analysis (Cox regression) with constant time.

We have used the data file <**Data_4.sav**> for the analysis. Our interest is to estimate the prevalence ratio of diabetes for males compared to females after controlling for age, family history of diabetes (variable name is "f_history") and peptic ulcer (variable name is "pepticulcer").

To analyze the data, first, we shall have to generate a constant time variable (say, c_time) that will have the same value (say, 1) for all the subjects (constant time). To generate the variable, use the following commands.

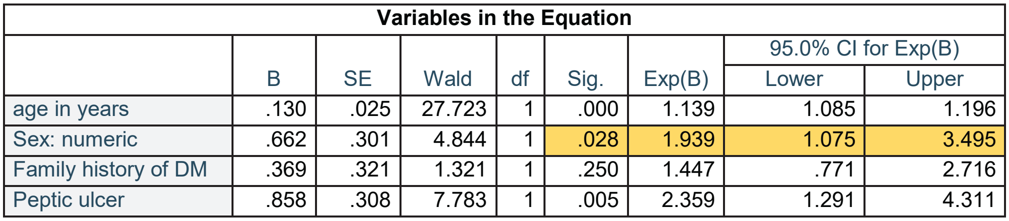
Transform > Compute variable > Type "c_time" in "Target variable" box > Click in the box under "Numeric expression" > Click 1 from the number pad > Ok (Fig 6.9)

This will generate the new variable "c_time" with all the values equal to 1. We shall use this variable (variable with constant time) for the analysis. To do the proportional hazards analysis with constant time, use the following commands.

Analyze > Survival > Cox Regression > Push "c_time" into the "Time" box > Push "diabetes" into the "Status" box > Click on "Define event" > In "Single value" box write 1 (since 1 is the code no. for "have diabetes") > Continue > Push "sex_1, age, f_hisotry and pepticulcer" into the "Covariate" box > Click on "Categorical" > Push "sex_1, f_history and pepticulcer" into the "Categorical covariates" box > Select "sex_1, f_history and pepticulcer" in the "Categorical covariates" box > Select "First" from "Reference category" under "Change contrast" (we are doing this to make "female", "no history of diabetes" and "no peptic ulcer" as the comparison group) > Change > Continue > Click on "Options" > Select "CI for Exp(B) and Correlation of estimates" > Continue > OK (Figs 17.1 to 17.4)

The above commands will produce several tables. We have provided only the relevant table (variables in the equation) needed for interpretation (Table 17.4). The column Exp(B) indicates the prevalence ratio. The table shows that the prevalence ratio of diabetes for males (since females are the comparison group) is 1.93 (95% CI: 1.07 to 3.49; p=0.028) after controlling for age, family history of diabetes and peptic ulcer.

Table 17.4 Variables in the Equation



Variables in the Equation

							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
age in years	.130	.025	27.723	1	.000	1.139	1.085	1.196
Sex: numeric	.662	.301	4.844	1	.028	1.939	1.075	3.495
Family history of DM	.369	.321	1.321	1	.250	1.447	.771	2.716
Peptic ulcer	.858	.308	7.783	1	.005	2.359	1.291	4.311

18

Non-parametric Methods

Non-parametric tests, in general, are done when the quantitative dependent variable is not normally distributed. Non-parametric tests are also used when the data are measured in nominal and ordinal scales. Box 18.1 shows the types of non-parametric methods recommended against the parametric tests, when the dependent variable is *not* normally distributed in the population. Note that non-parametric tests are less sensitive compared to the parametric tests and may, therefore, fail to detect differences between groups that actually exist. Use the data file <**Data_3.sav**> for practice.

Box 18.1 Types of non parametric techniques against the alternative parametric methods

Non-parametric test	Alternative parametric test
Mann-Whitney U test	Independent-samples t-test
Wilcoxon Signed Ranks test	Paired t-test
Kruskal-Wallis test	One-way ANOVA
Friedman test	One-way repeated measures ANOVA
Chi-square test for goodness-of-fit	None
Chi-square test of independence	None
Spearmen's correlation	Pearson's correlation

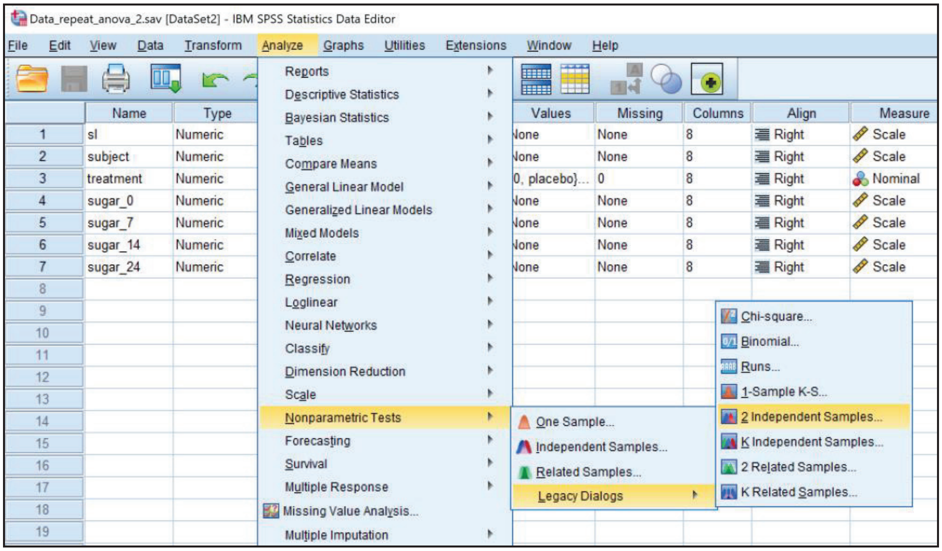
18.1 Mann-Whitney U test

This test is the alternative test for the Independent Samples t-test, when the dependent variable is not normally distributed. This test compares the differences between two groups on a continuous measure (variable). This test is based on the ranks of observations and is better than the median test. This test, tests the null hypothesis that the two population have equal medians. For example, we may want to know whether the median systolic BP (where the distribution of systolic BP is non-normal) of males and females is same.

18.1.1 Commands

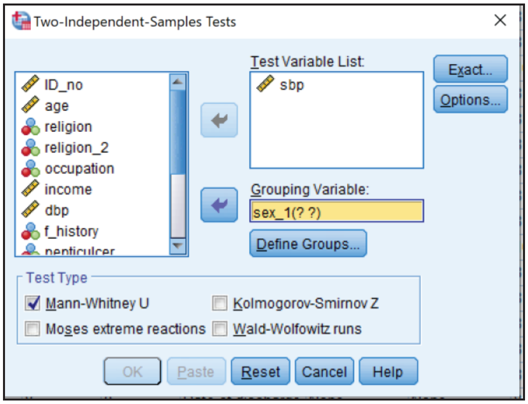
Analyze > Nonparametric tests > Legacy dialogs > 2 Independent samples > Select "sbp" and push into the "Test variable list" box > Select "sex_1" and push into the "Grouping variable" box > Click on "Define groups" > Write 0 in "Group1" box and 1 in "Group 2" box (note: our code nos. are 0 for female and 1 for male) > Continue > Select "Mann-Whitney" under "Test Type" > OK (Figs 18.1 and 18.2)

Figure 18.1



	Name	Type	Values	Missing	Columns	Align	Measure
1	sl	Numeric	lone	None	8	Right	Scale
2	subject	Numeric	lone	None	8	Right	Scale
3	treatment	Numeric	0, placebo}...	0	8	Right	Nominal
4	sugar_0	Numeric	lone	None	8	Right	Scale
5	sugar_7	Numeric	lone	None	8	Right	Scale
6	sugar_14	Numeric	lone	None	8	Right	Scale
7	sugar_24	Numeric	lone	None	8	Right	Scale
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							

Figure 18.2



Test Variable List: sbp

Grouping Variable: sex_1(? ?)

Test Type

☑ Mann-Whitney U ☐ Kolmogorov-Smirnov Z

☐ Moses extreme reactions ☐ Wald-Wolfowitz runs

18.1.2 Outputs

Table 18.1 Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
Systolic BP	210	113.75	123.00	142.00
Sex: numeric	210	.00	.00	1.00

Table 18.2 Mann-Whitney Ranks

	Sex: numeric	N	Mean Rank	Sum of Ranks
Systolic BP	Female	133	109.90	14616.50
	Male	77	97.90	7538.50
	Total	210		

18.3 Test Statistics[a]

	Systolic BP
Mann-Whitney U	4535.500
Wilcoxon W	7538.500
Z	-1.379
Asymp. Sig. (2-tailed)	.168

a. Grouping Variable: Sex: numeric

18.1.3 Interpretation

Our interest is in Table 18.3. Just look at the p-value (Asymp. Sig.) of the test. Here, the p-value is 0.168, which is >0.05. This indicates that the distribution of systolic BP among males and females is not different (or median systolic BP of males and females is not different). With this test result, the median systolic BP of females and males should be reported. To get the *median by sex*, use the following commands.

> Analyze > Compare means > Means > Select "sbp" and push it into the "Dependent list" box > Select "sex_1" and push it into the "Independent list" box under "Layer 1 of 1" > Options > Remove "Mean, Number of cases and Standard deviation" from the "Cell statistics" box > Select "Median" from "Statistics" box and push it into the "Cell statistics" box > Continue > OK

You will get the following tables (Tables 18.4 and 18.5). Table 18.5 shows the median systolic BP of males and females.

Table 18.4 Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Systolic BP * Sex: numeric	210	100.0%	0	0.0%	210	100.0%

Table 18.5 Median systolic BP by sex

Median	
Sex: numeric	Systolic BP
Female	124.00
Male	122.00
Total	123.00

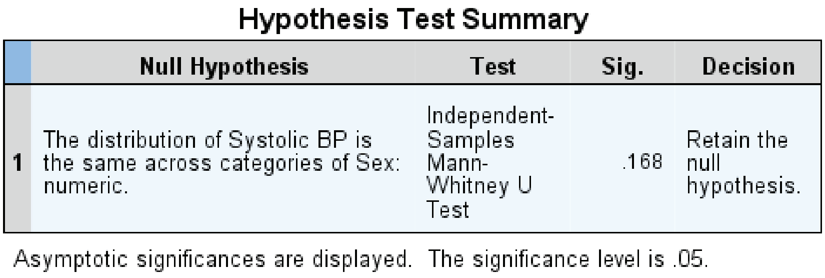
The Mann-Whitney test can also be done in a different way and will provide the same results. The alternative way of doing the Mann-Whitney test is to use the following commands.

> Analyze > Nonparametric tests > Independent samples > Select "Automatically compare distributions across groups" under "What is your objective" (it is usually the default) > Click on "Fields" (Fig 18.3) > Select "sbp" from "Fields" and push it into the "Test fields" box > Select "sex" and push it into the "Groups" box (Fig 18.4) > Run

Note: if you double click on the output table (Table 18.6) in SPSS program output, you will get the Figure 18.5.

The commands will provide you the following outputs (Table 18.6 and Fig 18.5).

Table 18.6 Mann-Whitney U Test results

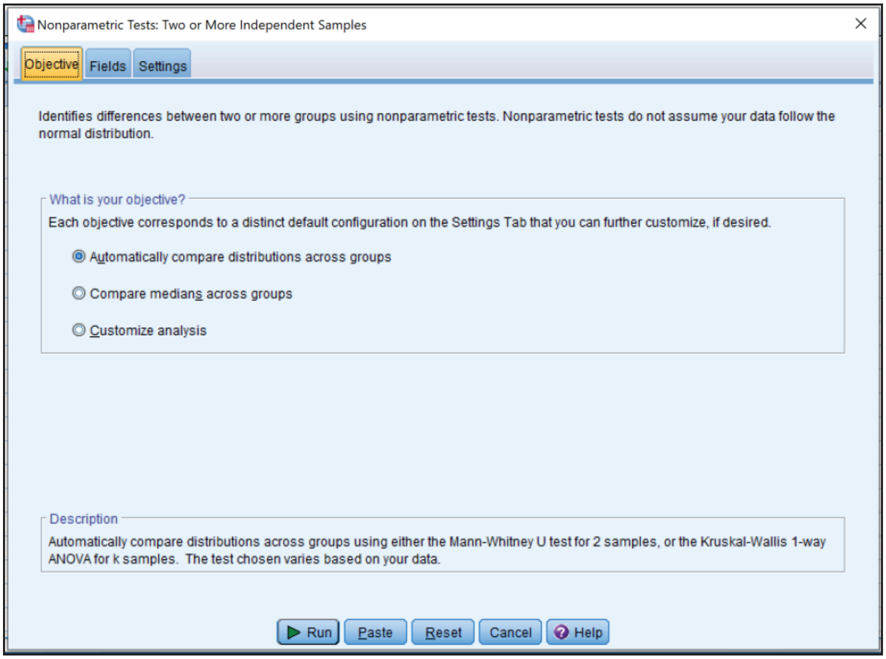


Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Systolic BP is the same across categories of Sex: numeric.	Independent-Samples Mann-Whitney U Test	.168	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 18.3



Identifies differences between two or more groups using nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

● Automatically compare distributions across groups

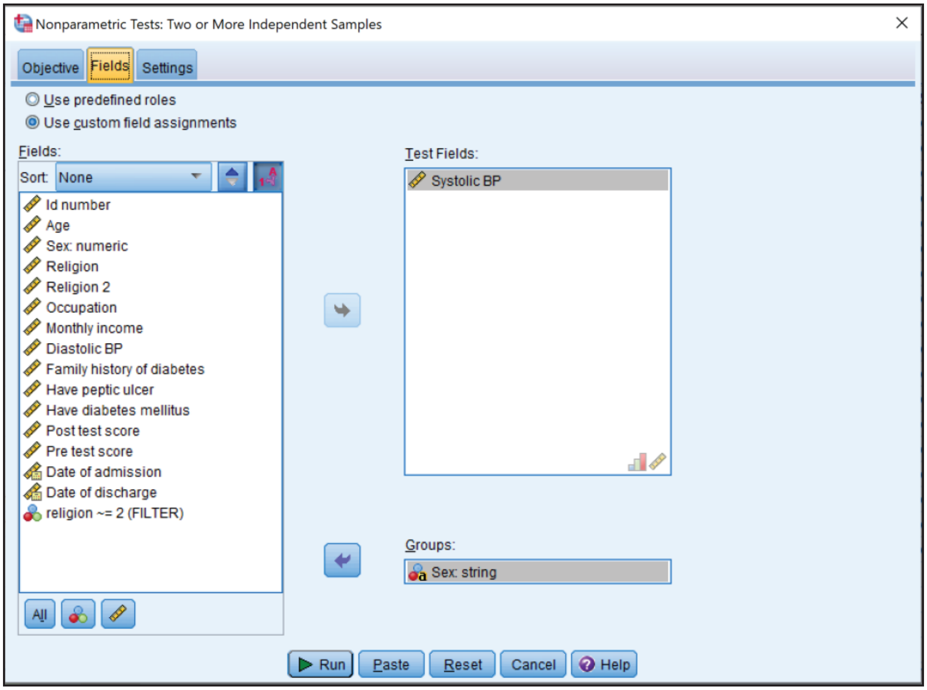
○ Compare medians across groups

○ Customize analysis

Description

Automatically compare distributions across groups using either the Mann-Whitney U test for 2 samples, or the Kruskal-Wallis 1-way ANOVA for k samples. The test chosen varies based on your data.

Figure 18.4



○ Use predefined roles

● Use custom field assignments

Fields:

Sort: None

- Id number
- Age
- Sex: numeric
- Religion
- Religion 2
- Occupation
- Monthly income
- Diastolic BP
- Family history of diabetes
- Have peptic ulcer
- Have diabetes mellitus
- Post test score
- Pre test score
- Date of admission
- Date of discharge
- religion ~= 2 (FILTER)

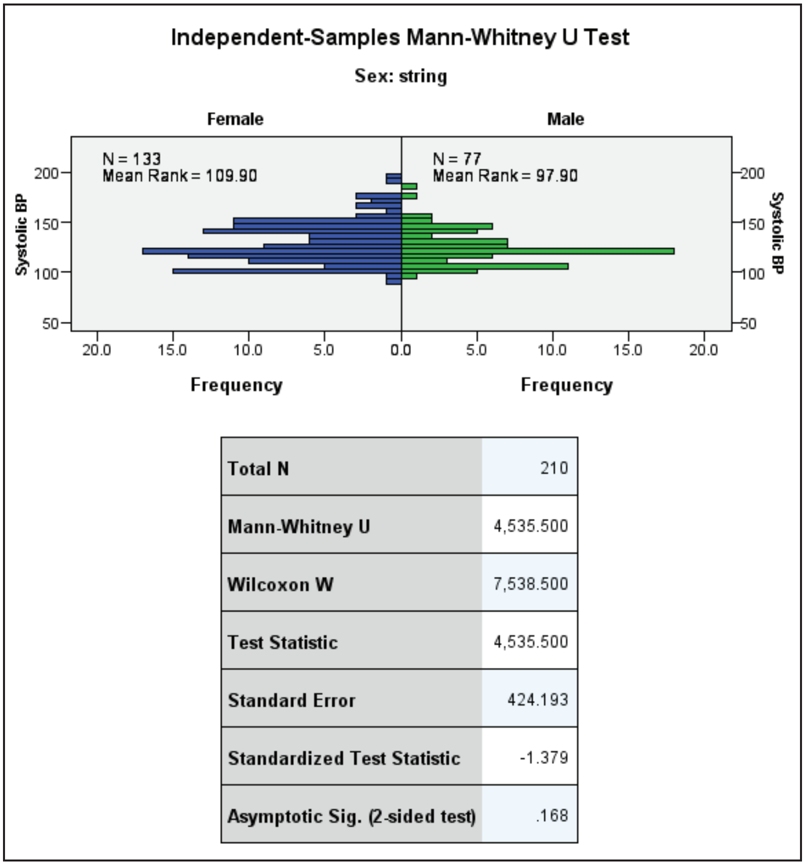
Test Fields:

- Systolic BP

Groups:

- Sex: string

Figure 18.5 Details about the Mann-Whitney test



Total N	210
Mann-Whitney U	4,535.500
Wilcoxon W	7,538.500
Test Statistic	4,535.500
Standard Error	424.193
Standardized Test Statistic	-1.379
Asymptotic Sig. (2-sided test)	.168

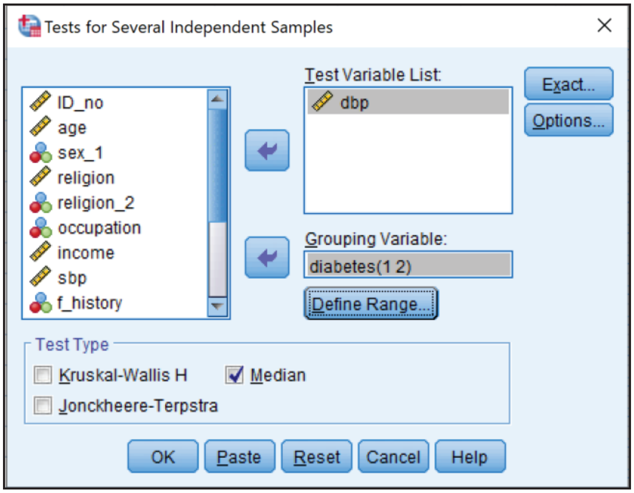
18.2 Median test

The median test is an alternative to the Mann-Whitney U test. Like the Mann-Whitney test, this test also compares the difference of medians between the two categories/groups on a continuous variable. This test is based on the number of observations below and above the common median. Suppose we want to determine whether the median diastolic BP of diabetics and non-diabetics is same in the population. Here, the null hypothesis is "the median diastolic BP of diabetics and non-diabetics is same in the population".

18.2.1 Commands

Analyze > Nonparametric tests > Legacy dialogs > K Independent samples > Select "dbp" and push it into the "Test variable list" box > Select "diabetes" and push it into the "Grouping variable" box > Click on "Define range" > Write 1 in "Minimum" box and 2 in "Maximum" box (note: our code nos. are 1 for have diabetes and 2 for do not have diabetes) > Continue > Options > Select "Quartile" > Continue > Deselect "Kruskal-Wallis H" and Select "Median" under "Test type" > OK (Fig 18.6)

Figure 18.6



Tests for Several Independent Samples

Test Variable List: dbp

Grouping Variable: diabetes(1 2)

Test Type

☐ Kruskal-Wallis H ☑ Median

☐ Jonckheere-Terpstra

18.2.2 Outputs

SPSS will provide the following outputs (Tables 18.7 to 18.9).

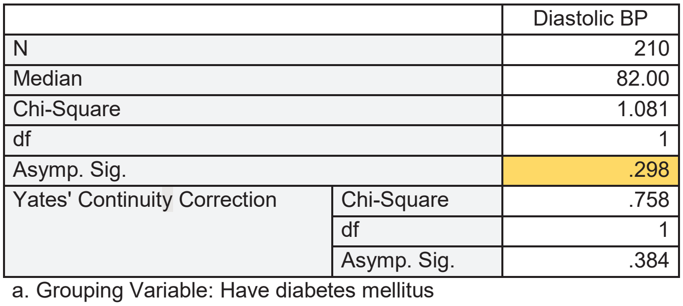
Table 18.7 Descriptive Statistics

		Percentiles		
	N	25th	50th (Median)	75th
Diastolic BP	210	74.00	82.00	90.00
Have diabetes mellitus	210	2.00	2.00	2.00

Table 18.8 Frequencies (below and above the common median)

		Have diabetes mellitus	
		Yes	No
Diastolic BP	> Median	23	70
	<= Median	22	95

Table 18.9 Test Statistics[a] (Median test results)



		Diastolic BP
N		210
Median		82.00
Chi-Square		1.081
df		1
Asymp. Sig.		.298
Yates' Continuity Correction	Chi-Square	.758
	df	1
	Asymp. Sig.	.384

a. Grouping Variable: Have diabetes mellitus

18.2.3 Interpretation

Table 18.7 shows the median diastolic BP (82.0) of the study subjects irrespective of having diabetes. Table 18.8 shows the frequencies of diabetic and non-diabetic persons above and below the median. For instance, 23 persons with diabetes have the diastolic BP above 82.0 mmHg (median) compared to 70 persons without diabetes. However, our interest is in Table 18.9. The Chi-square p-value of the test is 0.298, which is >0.05. We cannot, therefore, reject the null hypothesis. This indicates that the median diastolic BP of diabetic and non-diabetic persons in the population is not different. With these test results, one should report the median diastolic BP of diabetic and non-diabetic persons. To get the *median diastolic BP disaggregated by diabetes*, use the commands as shown in section 18.1.3.

18.3 Wilcoxon signed ranks test

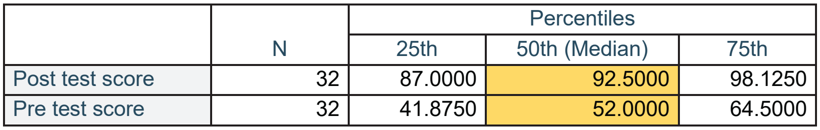
This test is the non-parametric alternative of the paired samples t-test. This test compares the distribution of two related samples (e.g., pre-test and post-test results). Wilcoxon test converts the scores into ranks and then compares. For example, in order to evaluate the impact of a training program, you have taken the pre- and post-tests, before and after the training. You want to assess if there is any change in the post-test scores compared to the pre-test scores due to the training.

18.3.1 Commands

Analyze > Nonparametric tests > Legacy dialogs > 2 Related Samples > Select “post-test and pre-test” and push them into the “Test Pairs” box > Select “Wilcoxon” under “Test type” (usually the default) > Options > Select “Quartile” > Continue > OK

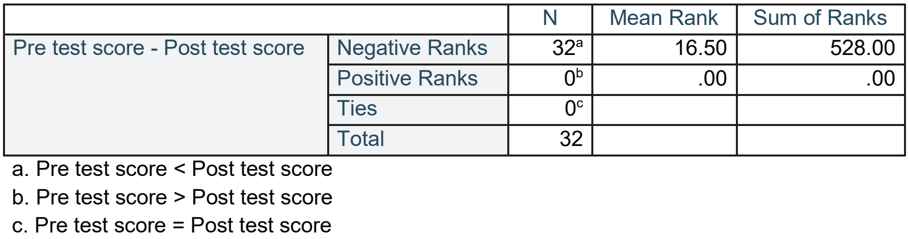
18.3.2 Outputs

Table 18.10 Descriptive Statistics



	N	Percentiles		
		25th	50th (Median)	75th
Post test score	32	87.0000	92.5000	98.1250
Pre test score	32	41.8750	52.0000	64.5000

Table 18.11 Ranks



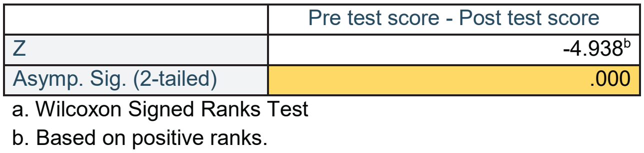
		N	Mean Rank	Sum of Ranks
Pre test score - Post test score	Negative Ranks	32[a]	16.50	528.00
	Positive Ranks	0[b]	.00	.00
	Ties	0[c]		
	Total	32		

a. Pre test score < Post test score

b. Pre test score > Post test score

c. Pre test score = Post test score

Table 18.12 Test Statistics[a]



	Pre test score - Post test score
Z	-4.938[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

18.3.3 Interpretation

Table 18.10 shows the descriptive statistics of pre- and post-test scores. The median (50th percentile) score of the pre-test is 52.0, while the median score is 92.5 for the post-test. The difference between these scores is quite big. Look at Table 18.12. The p-value of the Wilcoxon Signed Ranks test is 0.000, which is highly significant. This indicates that the pre- and post-test scores (medians) are significantly different. We, therefore, conclude that the training has significantly improved the knowledge of the participants (since the median of the post-test score is significantly higher than that of the pre-test score).

18.4 Kruskal-Wallis test

It is the non-parametric equivalent of the one-way ANOVA test. In this test, scores are converted into ranks and the mean rank of each group is compared. Suppose we want to test the hypothesis of whether the systolic BP is different among religious groups (Muslim, Hindu and Christian) [the null hypothesis is “systolic BP is not different across the religious groups”].

18.4.1 Commands

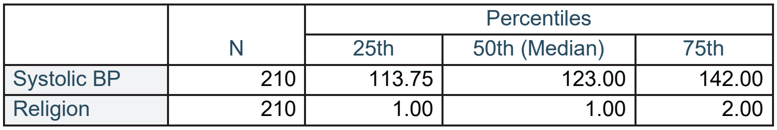
Analyze > Nonparametric tests > Legacy dialogs > K Independent Samples > Select “sbp” and push it into “Test variable list” box > Select “religion” and push it into “Grouping variable” box > Click “Define range” > Write 1 in “Minimum” box and 3 in “Maximum” box (the religion has code numbers from 1 to 3) > Continue > Options > Select “Quartile” > Continue > Select “Kruskal Wallis H” under “Test type” (usually the default) > OK

To get the median of systolic BP by religious groups, use the following commands.

Analyze > Compare means > Means > Select “sbp” and push into the “Dependent list” box > Select “religion” and push it into the “Independent list” box under “Layer 1 of 1” > Options > Remove “Mean, Number of cases and Standard deviation” from the “Cell statistics” box > Select “Median” from “Statistics” box and push it into the “Cell statistics” box > Continue > OK (Table 18.16)

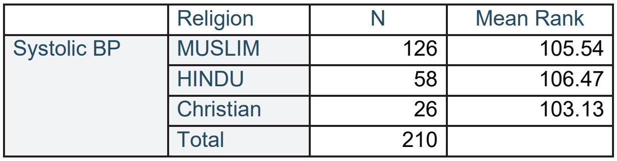
18.4.2 Outputs

Table 18.13 Descriptive Statistics



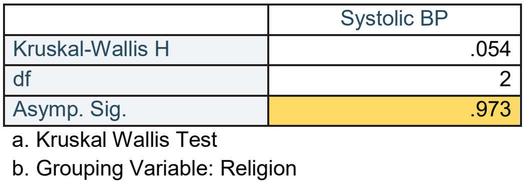
		Percentiles		
	N	25th	50th (Median)	75th
Systolic BP	210	113.75	123.00	142.00
Religion	210	1.00	1.00	2.00

Table 18.14 Ranks



	Religion	N	Mean Rank
Systolic BP	MUSLIM	126	105.54
	HINDU	58	106.47
	Christian	26	103.13
	Total	210	

Table 18.15 Test Statistics[a,b]

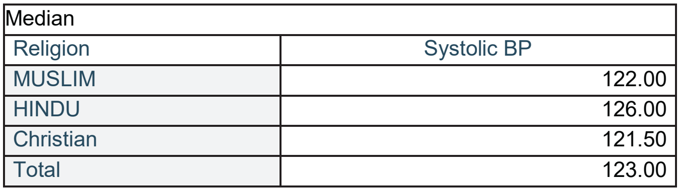


	Systolic BP
Kruskal-Wallis H	.054
df	2
Asymp. Sig.	.973

a. Kruskal Wallis Test

b. Grouping Variable: Religion

Table 18.16 Median of systolic BP in different religious groups



Median	
Religion	Systolic BP
MUSLIM	122.00
HINDU	126.00
Christian	121.50
Total	123.00

18.4.3 Interpretation

Table 18.15 shows the Kruskal-Wallis test results (dependent variable is the systolic BP and the grouping variable is religion with 3 levels – Muslim, Hindu and Christian as shown in Table 18.14). The p-value (Asymp. Sig.) of the Chi-square test is 0.973, which is >0.05. Therefore, we are unable to reject the null hypothesis. We conclude that the median systolic BP among the religious groups is not significantly different. The median systolic BP in different religious groups is provided in Table 18.16.

18.5 Friedman test

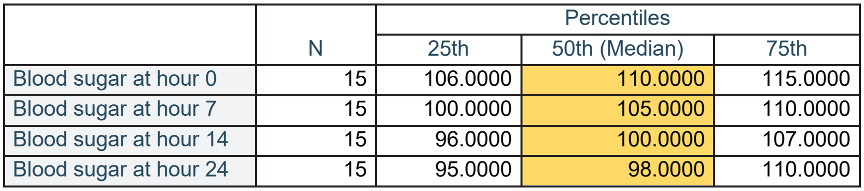
The Friedman test is the non-parametric alternative of the one-way repeated measures ANOVA test. For example, we are interested to evaluate the changes in blood sugar levels (if they are different or not) at four different time intervals (e.g., at hour 0, hour 7, hour 14 and hour 24) after administration of a drug. To conduct this study, we have selected 15 individuals randomly from a population and measured their blood sugar levels at the baseline (hour 0). All the individuals were then given the drug, and their blood sugar levels were measured again at hour 7, hour 14 and hour 24. The blood sugar levels at hour 0, hour 7, hour 14, and hour 24 are named in SPSS as Sugar_0, Sugar_7, Sugar_14 and Sugar_24, respectively. Use the data file <**Data_Repeat_anova_2.sav**> for practice.

18.5.1 Commands

> Analyze > Nonparametric Tests > Legacy dialogs > K Related Samples > Select "sugar_0, sugar_7, sugar_14 and sugar_24" and push them into "Test variables" box > Statistics > Select "Quartile" > Continue > Select "Friedman" under "Test type" > OK

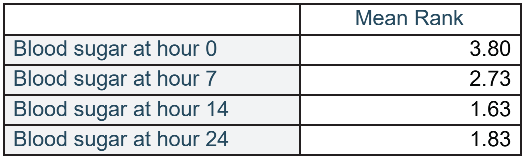
18.5.2 Outputs

Table 18.17 Descriptive Statistics



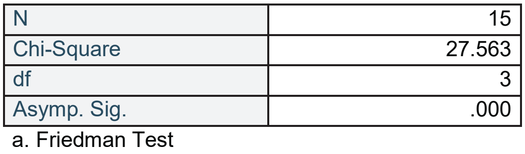
	N	Percentiles		
		25th	50th (Median)	75th
Blood sugar at hour 0	15	106.0000	110.0000	115.0000
Blood sugar at hour 7	15	100.0000	105.0000	110.0000
Blood sugar at hour 14	15	96.0000	100.0000	107.0000
Blood sugar at hour 24	15	95.0000	98.0000	110.0000

Table 18.18 Ranks



	Mean Rank
Blood sugar at hour 0	3.80
Blood sugar at hour 7	2.73
Blood sugar at hour 14	1.63
Blood sugar at hour 24	1.83

Table 18.19 Test Statistics[a]



N	15
Chi-Square	27.563
df	3
Asymp. Sig.	.000

a. Friedman Test

18.5.3 Interpretation

Outputs are provided in Tables 18.17 to 18.19. Table 18.17 shows the median blood sugar levels at 4 different time periods. Look at the Friedman test results provided in Table 18.19. The Chi-square value is 27.56 and the p-value (Asymp. Sig.) is 0.000, which is <0.05. This indicates that there is a significant difference in blood sugar levels across the 4 time periods (p<0.001). The findings suggest that the drug is effective in reducing blood sugar levels, since the median blood sugar levels have reduced over time (Table 18.7).

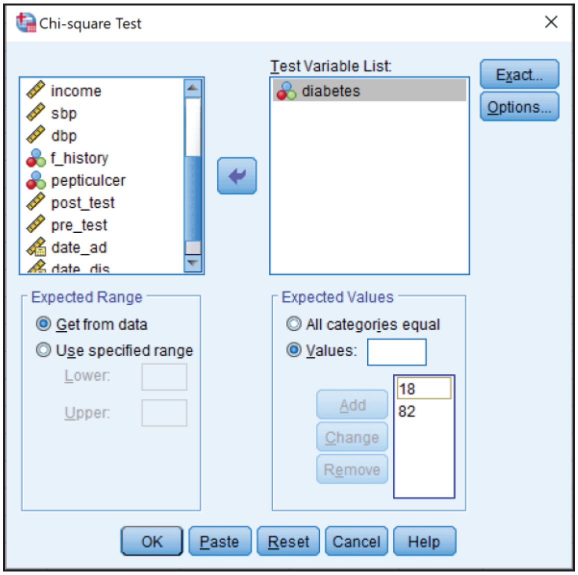
18.6 Chi-square test for goodness-of-fit

The Chi-square test of independence, which is the most frequently used test to determine the association between two categorical variables, has been discussed in chapter 11. The Chi-square test for goodness-of-fit is also referred to as one-sample Chi-square test. It is often used to compare the proportion of cases with a hypothetical proportion. For instance, we have conducted a survey taking a random sample from a population, and the data shows that the prevalence of diabetes is 21.4%. Now we want to test the hypothesis whether the prevalence of diabetes in the population is 18% or not (the null hypothesis is “the prevalence of diabetes in the population is 18%”). To have the answer, we shall do the Chi-square test for goodness-of-fit (seldom we test such a hypothesis).

18.6.1 Commands

Analyze > Nonparametric tests > Legacy dialogs > Chi-square > Move the variable “diabetes” into the “Test variable list” box > Select “Values” under “Expected values” > Write “18” in the box > Add > Again write “82” (100 minus 18) in the box after deleting 18 > Add > OK (Fig 18.7)

Figure 18.7



Chi-square Test

Test Variable List: diabetes

Expected Range
● Get from data
○ Use specified range
Lower:
Upper:

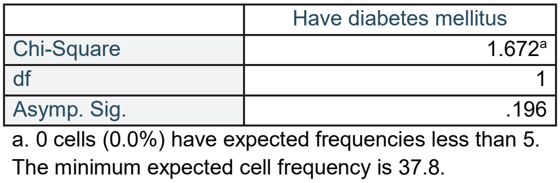
Expected Values
○ All categories equal
● Values:
18
82

18.6.2 Outputs

Table 18.20 Have diabetes mellitus

	Observed N	Expected N	Residual
Yes	45	37.8	7.2
No	165	172.2	-7.2
Total	210		

Table 18.21 Test Statistics



	Have diabetes mellitus
Chi-Square	1.672[a]
df	1
Asymp. Sig.	.196

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 37.8.

18.6.3 Interpretation

Table 18.20 provides the observed and expected frequencies for those who have diabetes (as "Yes") and those who do not have diabetes (as "No"). These are the descriptive information that you may not need to report. Table 18.21 is the main table to interpret the results. Our interest is at the p-value. The Chi-square value, as shown in the table, is 1.672 and the p-value is 0.196. Since the p-value is >0.05, we cannot reject the null hypothesis. This indicates that the prevalence of diabetes in the population may not be different from 18%.

19

Checking Reliability of Scales: Cronbach's Alpha

When researchers select a scale (e.g., a scale to measure depression) in their study, it is important to check that the scale is reliable. One of the ways to check the internal consistency (reliability) of a scale is to calculate the Cronbach's alpha coefficient. Cronbach's alpha indicates the degree to which the items that make up the scale correlate with each other in the group.

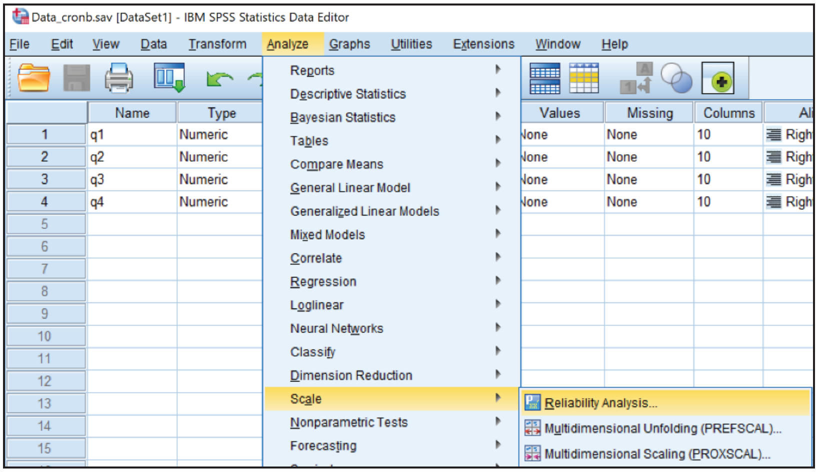
Ideally, the Cronbach's alpha value should be above 0.7. However, this value is sensitive to the number of items on the scale. If the number of items on the scale is less than 10, the Cronbach's alpha coefficient tends to be low. In such a situation, it is appropriate to use the "*mean inter-item correlations*". The optimum range of the mean inter-item correlation is between 0.2 and 0.4. Use the data file <**Data_cronb.sav**> for practice.

19.1 Cronbach's alpha

Before using the procedure to get the Cronbach's alpha coefficient, be sure that all the negatively worded values are "reversed" by recoding. If this is not done, it will produce a very low (or negative) value of the Cronbach's alpha coefficient. Assume that a researcher has used a scale to identify/measure depression. The scale has 4 questions, q1, q2, q3 and q4. To get the Cronbach's alpha coefficient, use the following commands.

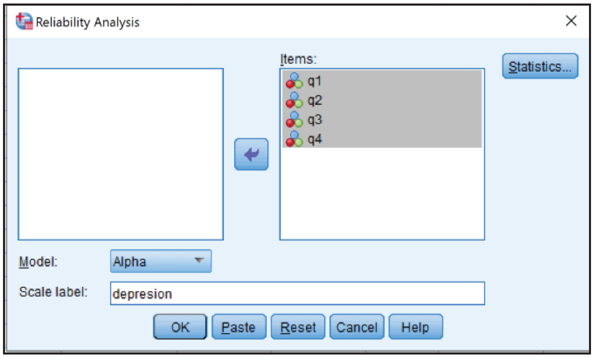
> Analyze > Scale > Reliability analysis > Select all the items (q1, q2, q3 & q4) that construct the scale and push them into the "Items" box > Make sure that "Alpha" is selected in "Model" section (usually the default) > Type name of the scale (e.g., depression or give any other name suitable for the data) in the "Scale label" box > Statistics > Select "Item, Scale, and Scale if item deleted" under "Descriptives for" section > Select "Correlations" under "Inter-item" section > Select "Correlations" under "Summaries" section > Continue > OK (Figs 19.1 to 19.3)

Figure 19.1


Data_cronb.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Reliability Analysis...
Multidimensional Unfolding (PREFSCAL)...
Multidimensional Scaling (PROXSCAL)...

	Name	Type	Values	Missing	Columns	Al
1	q1	Numeric	None	None	10	Righ
2	q2	Numeric	None	None	10	Righ
3	q3	Numeric	None	None	10	Righ
4	q4	Numeric	None	None	10	Righ

Figure 19.2

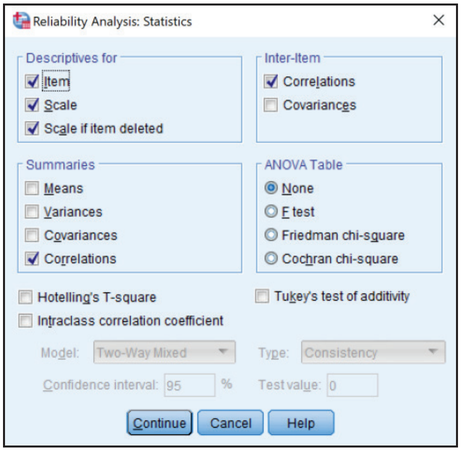


Items:
q1
q2
q3
q4

Model: Alpha

Scale label: depression

Figure 19.3



Descriptives for
☑ Item
☑ Scale
☑ Scale if item deleted

Inter-Item
☑ Correlations
□ Covariances

Summaries
□ Means
□ Variances
□ Covariances
☑ Correlations

ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

□ Hotelling's T-square
□ Tukey's test of additivity
□ Intraclass correlation coefficient

Model: Two-Way Mixed
Type: Consistency
Confidence interval: 95 %
Test value: 0

19.1.1 Outputs

Table 19.1 Case Processing Summary

		N	%
Cases	Valid	60	100.0
	Excluded[a]	0	.0
	Total	60	100.0

a. Listwise deletion based on all variables in the procedure.

Table 19.2 Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.839	.840	4

Table 19.3 Item Statistics

	Mean	Std. Deviation	N
q1	3.1333	1.08091	60
q2	3.2667	1.00620	60
q3	3.0167	1.08130	60
q4	3.2833	1.13633	60

Table 19.4 Inter-Item Correlation Matrix

	q1	q2	q3	q4
q1	1.000	.512	.491	.548
q2	.512	1.000	.635	.630
q3	.491	.635	1.000	.589
q4	.548	.630	.589	1.000

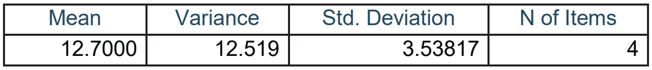
Table 19.5 Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Inter-Item Correlations	.567	.491	.635	.143	1.292	.003	4

Table 19.6 Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
q1	9.5667	7.741	.600	.364	.827
q2	9.4333	7.572	.711	.518	.781
q3	9.6833	7.373	.677	.476	.794
q4	9.4167	6.993	.705	.500	.781

Table 19.7 Scale Statistics



Mean	Variance	Std. Deviation	N of Items
12.7000	12.519	3.53817	4

19.1.2 Interpretation

The Reliability Statistics table (Table 19.2) shows the Cronbach's alpha value. In this example, the value is 0.839. This indicates a very good correlation among the items on the scale (i.e., the scale is reliable).

However, before looking at the value of Cronbach's alpha, look at the table "Inter-item Correlation Matrix" (Table 19.4). All the values in the table should be *positive* (all are positive in our example). One or more negative values (if there is any) indicate that some of the items have not been "reverse scored" correctly. This information is also provided in the table "Item-total Statistics (Table 19.6)". All the values under "Corrected item - Total Correlation" should be positive (there should not have any negative value).

The corrected item-total correlation (Table 19.6) indicates the degree to which each item correlates with the total score. In our example, the values are 0.60, 0.71, 0.67 and 0.70. A small value for any item (<0.3) could be a problem. If the Cronbach's alpha value (Table 19.2) and the corrected item-total correlation value for any item is small (<0.7 and <0.3, respectively), one may consider omitting the item from the scale that has a small value. In our example, there is no such problem.

However, if the number of items is small on the scale (fewer than 10), it may be difficult to get a reasonable Cronbach's alpha value. In such a situation, report the *Mean Inter-item Correlation* value (Summary-item Statistics table; Table 19.5). In this example, the Inter-item Correlation values range from 0.491 to 0.635, and the mean is 0.567 (optimum range of the mean is 0.2 to 0.4). This indicates a strong relationship among the items.

20

Analysis of Covariance (ANCOVA)

ANCOVA stands for Analysis of Covariance, which is done to statistically control the extraneous variable(s) [called covariate] for the comparison of means of two or more groups. It is similar to ANOVA. In ANOVA, one can incorporate only the categorical independent variables to have the main effect and interaction. But in ANCOVA, one can incorporate both the categorical and quantitative variables in the model, including the interaction between categorical and quantitative independent variables. The ANCOVA can be performed as One-way, Two-way and Multivariate ANCOVA techniques. Use the data file <**Data_3.sav**> for practice.

20.1 One-way ANCOVA

The purpose of doing the one-way ANCOVA test is to assess the differences in means of the dependent variable (e.g., systolic BP) against a categorical variable (e.g., sex, or effect of drugs) after controlling for the quantitative variable(s) [called covariates, such as age, diastolic BP] in the model. The one-way ANCOVA test involves at least three variables:

- One quantitative *dependent* variable (e.g., systolic BP, post-test score, blood sugar level);
- Only one categorical *independent* variable with two or more levels (e.g., sex, type of intervention, or type of drug); and
- One (or more) *covariate* (continuous quantitative variable), e.g., diastolic BP, age, pre-test score, baseline blood sugar level.

The covariates to be selected for the model should be one or more continuous variables and they should significantly correlate with the dependent variable. One can also include categorical variables as covariates in the model.

Suppose the researcher is interested in comparing the effectiveness of 3 drugs (drug A, drug B and drug C) in reducing the systolic BP. To conduct the study, the researcher has randomly selected three groups of people and assigned these drugs, one in each group. In this scenario, one-way ANOVA could be used. However, it was observed that the mean age and pre-treatment systolic BP of these three groups are not the same. Since age and pre-treatment systolic BP can influence the effectiveness of the drugs in reducing the systolic BP, it requires adjustment for these variables (age and pre-treatment systolic BP are the covariates) to conclude the results. In such a situation, one-way ANCOVA can be used. Note that for ANCOVA, the independent variable must be a categorical variable (here it is “type of drug”). ANCOVA can adjust for more than one covariate, either continuous or categorical.

Another example: Assume that you have organized a training program. To evaluate the effectiveness of the training, you have taken the pre- and post-tests of the participants. Now, you want to conclude if males and females (independent variable) have similar performance in the post-test (dependent variable), after controlling for age and pre-test results (covariates). One-way ANCOVA is the appropriate test for both these situations, if the assumptions are met.

Hypothesis

Assume that you want to assess if the mean systolic BP (dependent variable) is same among males and females (independent variable) after controlling for diastolic BP (covariate).

H_0: There is no difference in the mean systolic BP between males and females in the population (after controlling for diastolic BP).

H_A: The mean systolic BP of males and females is different in the population.

Assumptions

1. The dependent variable is normally distributed at each level of the independent variable;
2. The variances of the dependent variable for each level of the independent variab- -le are same (homogeneity of variances);
3. The covariates (if more than one) are not strongly correlated with each other (r<0.8);
4. There is a linear relationship between the dependent variable and the covariates at each level of the independent variable;
5. There is no interaction between the covariate (diastolic BP) and the independent variable (sex) [called *homogeneity of regression slopes*].

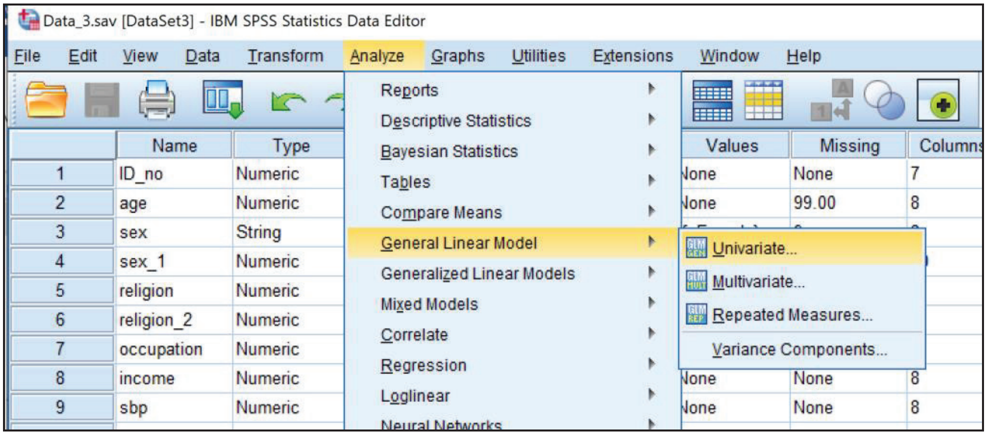
20.1.1 Commands

A. Commands: Homogeneity of regression slopes (Assumption 5)

First, we shall have to check the *homogeneity of regression slopes*, using the following commands. Note that, the SPSS variable names for sex is "sex_1 (0= female; 1= male)", Systolic BP is "sbp" and Diastolic BP is "dbp".

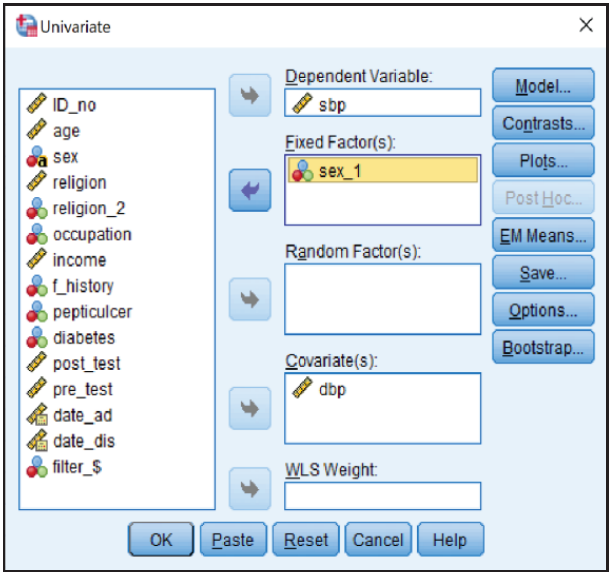
> Analyze > General linear model > Univariate > Push "sbp" into the "Dependent variables" box > Push "sex_1" into the "Fixed factor" box > Push "dbp" in the "Covariate" box > Click "Model" > Select "Build terms" under "Specify model" > Confirm that interaction option is showing in the "Build terms" box > Push "sex_1" and "dbp" into the "Model" box > Click on "sex_1" in "Factors & Covariates" box > Pressing the control button click on the "dbp" in "Factors & Covariates" box > Push them into the "Model" box (you will see "dbp*sex_1" in the Model box) > Continue > OK (Figs 20.1 to 20.3)

Figure 20.1



	Name	Type	Values	Missing	Columns
1	ID_no	Numeric	None	None	7
2	age	Numeric	None	99.00	8
3	sex	String			
4	sex_1	Numeric			
5	religion	Numeric			
6	religion_2	Numeric			
7	occupation	Numeric			
8	income	Numeric	None	None	8
9	sbp	Numeric	None	None	8

Figure 20.2



Dependent Variable: sbp

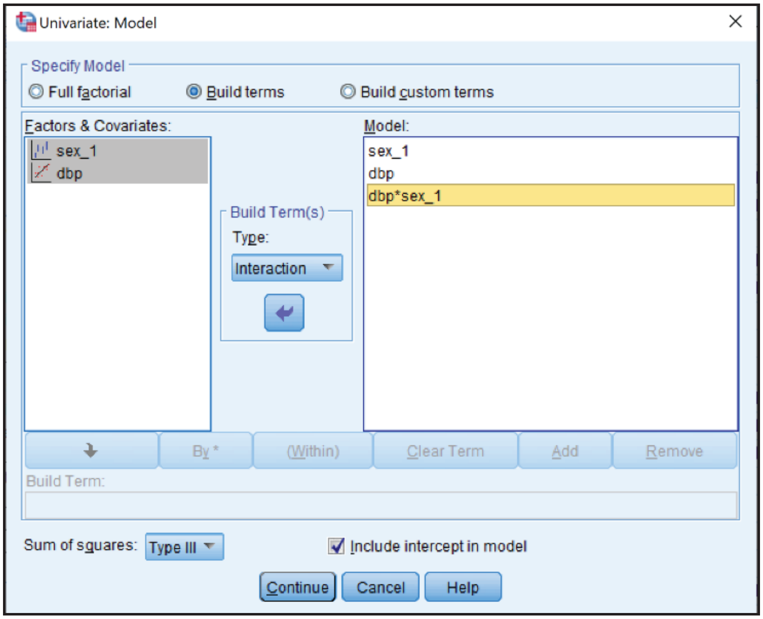
Fixed Factor(s): sex_1

Random Factor(s):

Covariate(s): dbp

WLS Weight:

Figure 20.3



Specify Model

○ Full factorial ● Build terms ○ Build custom terms

Factors & Covariates: sex_1, dbp

Model:

sex_1
dbp
dbp*sex_1

Type: Interaction

Build Term:

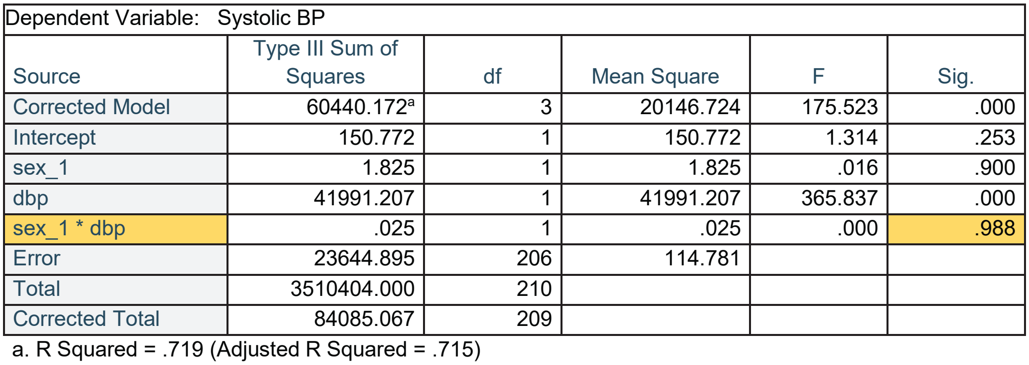
Sum of squares: Type III

☑ Include intercept in model

20.1.2 Outputs: Homogeneity of regression slopes

SPSS will provide two tables, but our interest is in table "Tests of between-subjects effects" (Table 20.1).

Table 20.1 Tests of Between-Subjects Effects



Dependent Variable: Systolic BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	60440.172[a]	3	20146.724	175.523	.000
Intercept	150.772	1	150.772	1.314	.253
sex_1	1.825	1	1.825	.016	.900
dbp	41991.207	1	41991.207	365.837	.000
sex_1 * dbp	.025	1	.025	.000	.988
Error	23644.895	206	114.781		
Total	3510404.000	210			
Corrected Total	84085.067	209			

a. R Squared = .719 (Adjusted R Squared = .715)

20.1.3 Interpretation: Homogeneity of regression slopes

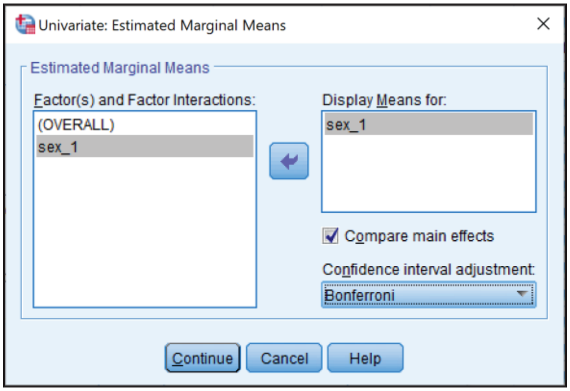
Look at Table 20.1 (Tests of between-subjects effects). Our interest is in the significance level (Sig.) of the interaction (Sex_1*dbp). We can see that the p-value for interaction is 0.988, which is >0.05. This indicates that the *homogeneity of regression slopes* assumption is not violated. A p-value of <0.05 indicates that the regression slopes are not homogeneous and the ANCOVA test is inappropriate.

B. Commands: One-way ANCOVA

To perform the one-way ANCOVA, use the following commands.

> Analyze > General linear model > Univariate > Push "sbp" into the "Dependent variables" box > Push "sex_1" into the "Fixed factor" box > Push "dbp" in the "Covariate" box > Click "Model" > Select "Full factorial" under "Specify model" > Continue > Click on "EM Means" > Select "sex_1" and push it into the "Display means for" box (this would provide the adjusted means) > Select "Compare main effects" > Select "Bonferroni" from "Confidence interval adjustment" (Fig 20.4) > Continue > Options > Select "Descriptive statistics, Estimates of effect size, and Homogeneity tests" under "Display" section > Continue > OK

Figure 20.4


Univariate: Estimated Marginal Means

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)
sex_1

Display Means for:
sex_1

☑ Compare main effects

Confidence interval adjustment: Bonferroni

20.1.4 Outputs: One-way ANCOVA

Table 20.2 Between-Subjects Factors

		Value Label	N
Sex: numeric	0	Female	133
	1	Male	77

Table 20.3 Descriptive Statistics

Dependent Variable: Systolic BP

Sex: numeric	Mean	Std. Deviation	N
Female	129.57	21.377	133
Male	124.56	17.221	77
Total	127.73	20.058	210

Table 20.4 Levene's Test of Equality of Error Variances[a]

Dependent Variable: Systolic BP

F	df1	df2	Sig.
.147	1	208	.702

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + dbp + sex_1

Table 20.5 Tests of Between-Subjects Effects

Dependent Variable: Systolic BP						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	60440.147[a]	2	30220.074	264.562	.000	.719
Intercept	208.045	1	208.045	1.821	.179	.009
dbp	59214.639	1	59214.639	518.396	.000	.715
sex_1	144.111	1	144.111	1.262	.263	.006
Error	23644.919	207	114.227			
Total	3510404.000	210				
Corrected Total	84085.067	209				

a. R Squared = .719 (Adjusted R Squared = .716)

Table 20.6 Estimates (estimated marginal)

Dependent Variable: Systolic BP				
			95% Confidence Interval	
Sex: numeric	Mean	Std. Error	Lower Bound	Upper Bound
Female	127.091[a]	.933	125.252	128.931
Male	128.842[a]	1.232	126.413	131.272

a. Covariates appearing in the model are evaluated at the following values: Diastolic BP = 82.77.

Table 20.7 Pairwise Comparisons

Dependent Variable: Systolic BP						
					95% Confidence Interval for Difference[a]	
(I) Sex: numeric	(J) Sex: numeric	Mean Difference (I-J)	Std. Error	Sig.[a]	Lower Bound	Upper Bound
Female	Male	-1.751	1.559	.263	-4.825	1.322
Male	Female	1.751	1.559	.263	-1.322	4.825

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Table 20.8 Pairwise Comparisons

Dependent Variable: Systolic BP						
					95% Confidence Interval for Difference[a]	
(I) Religion	(J) Religion	Mean Difference (I-J)	Std. Error	Sig.[a]	Lower Bound	Upper Bound
MUSLIM	HINDU	-.448	1.705	1.000	-4.562	3.666
	Christian	.948	2.313	1.000	-4.635	6.532
HINDU	MUSLIM	.448	1.705	1.000	-3.666	4.562
	Christian	1.397	2.535	1.000	-4.721	7.514
Christian	MUSLIM	-.948	2.313	1.000	-6.532	4.635
	HINDU	-1.397	2.535	1.000	-7.514	4.721

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

20.1.5 Interpretation: One-way ANCOVA

Tables 20.2 and 20.3 display the descriptive statistics. Table 20.3 shows the *unadjusted* means of systolic BP by sex (female: 129.5 and male: 124.5).

Table 20.4 shows the Levene's test of Equality of Error Variances. This is the test for assumption 2. We expect the p-value (Sig.) to be >0.05 to meet the assumption. In this example, the p-value is 0.70, which is more than 0.05. This means that the variances of the dependent variable (systolic BP) are equal at each level of the independent variable (sex).

Table 20.5 (tests of between-subjects effects) is the main table showing the results of the one-way ANCOVA test. We tested the hypothesis of whether the population mean of systolic BP in males and females is the same after *controlling for diastolic BP*. Look at the p-value for sex (sex_1) in the table, and it is 0.263. Since the p-value is >0.05, we cannot reject the null hypothesis. This indicates that the mean systolic BP (in the population) of males and females is not different after controlling for diastolic BP. Also look at the value for Partial Eta Squared. Eta indicates the amount of variance (also called effect size) in the dependent variable that is explained by the independent variable (sex). We can see that the effect size is very small (0.006 or 0.6%).

We can also assess the influence of covariate (diastolic BP) on the dependent variable (systolic BP). The p-value for diastolic BP is 0.000, which is highly significant. This indicates that there is a significant association between systolic and diastolic BP, after *controlling for sex*. The value of the Partial Eta Squared for diastolic BP is 0.715 (71.5%). This means that 71.5% variance of systolic BP can be explained by the diastolic BP after controlling for sex.

Table 20.6 (estimated marginal) shows the adjusted (adjusted for diastolic BP) means of the dependent variable (systolic BP) at different levels of the independent variable (sex). We can see that the mean systolic BP of females is 127.09 mmHg and that of the males is 128.84 mmHg after adjusting for diastolic BP (note that the adjusted means are different from the unadjusted means as shown in Table 20.3).

Table 20.7 is the table for pairwise comparison. This table is not necessary in this example, since the independent variable (sex) has two levels. If the independent variable has more than two levels, then the table for pairwise comparison is important to look at, especially if there is a significant association between the dependent and independent variables. Look at Table 20.8 [this is an additional table we have provided where the independent variable (religion) has three categories], which shows the pairwise comparison of mean systolic BP by religious groups. The results indicate that there is no significant difference in mean systolic BP among different religious groups, after controlling for diastolic BP since all the p-values are >0.05.

20.2 Two-way ANCOVA

In two-way ANCOVA, there are two independent categorical variables with two or more levels/categories, while in one-way ANCOVA, there is only one independent categorical variable with two or more levels. Therefore, in two-way ANCOVA, four variables are involved. They are:

- One continuous *dependent* variable (e.g., diastolic BP, blood sugar, post-test score);
- Two categorical *independent* variables (with two or more levels) [e.g., occupation, diabetes, type of drug]; and
- One or more continuous *covariates* (e.g., age, systolic BP, income).

The two-way ANCOVA provides information, *after controlling for the covariate(s)*, on:

1. Whether there is a significant main effect of the first independent variable (e.g., occupation) on the dependent variable;
2. Whether there is a significant main effect of the second independent variable (e.g., diabetes) on the dependent variable; and
3. Whether there is an interaction between the independent variables (e.g., occupation and diabetes).

Suppose we want to assess, after controlling for age (covariate):

- Whether occupation influences the diastolic BP (i.e., is mean diastolic BP in different occupational groups same?);
- Whether diabetes influences the diastolic BP (i.e., is the mean diastolic BP same for diabetics and non-diabetics?); and
- Does the influence of occupation on diastolic BP depend on the presence of diabetes (i.e., is there interaction between occupation and diabetes)?

Questions 1 and 2 refer to the *main effect*, while question 3 explains the interaction of two independent variables (occupation and diabetes) on the dependent variable (diastolic BP). For the analysis, we shall use the data file <**Data_3.sav**>. Note that the SPSS variable names of diastolic BP is “dbp”, occupation is “occupation”, diabetes is “diabetes” and age is “age”.

Assumptions

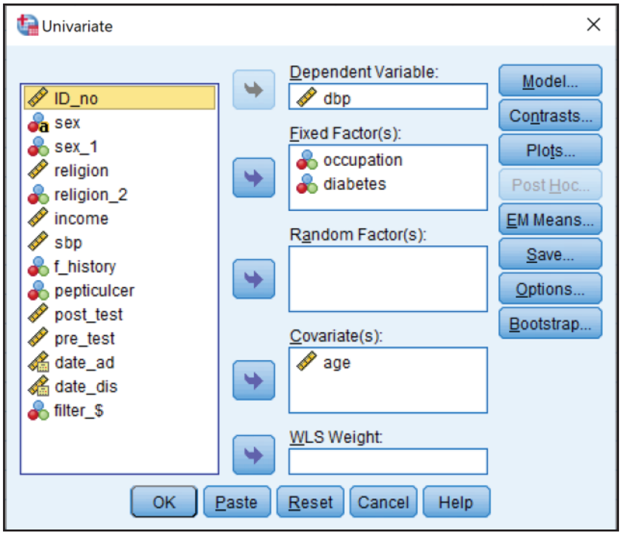
All the assumptions mentioned under one-way ANCOVA are applicable for two-way ANCOVA. Look at one-way ANCOVA for the assumptions and how to check the Homogeneity of regression slopes.

20.2.1 Commands

To perform the two-way ANCOVA, use the following commands.

> Analyze > General linear model > Univariate > Push "dbp" into the "Dependent variable" box > Push "occupation" and "diabetes" into the "Fixed factor" box > Push "age" into the "Covariate" box (Fig 20.5) > Click "Model" > Select "Full Factorial" > Continue > Click on "EM Means" > Select "occupation, diabetes and occupation*diabetes" and push them into the "Display means for" box (this would provide the adjusted means of the diastolic BP for occupation and diabetes) > Select "Compare main effects" > Select "Bonferroni" from "Confidence interval adjustment" (Fig 20.6) > Options > Select "Descriptive statistics" and "Homogeneity tests" > Continue > Plots > Select "occupation" and push it into the "Horizontal" box > Select "diabetes" and push it into the "Separate lines" box > Click "Add" > Continue > OK

Figure 20.5



Univariate

Dependent Variable: dbp

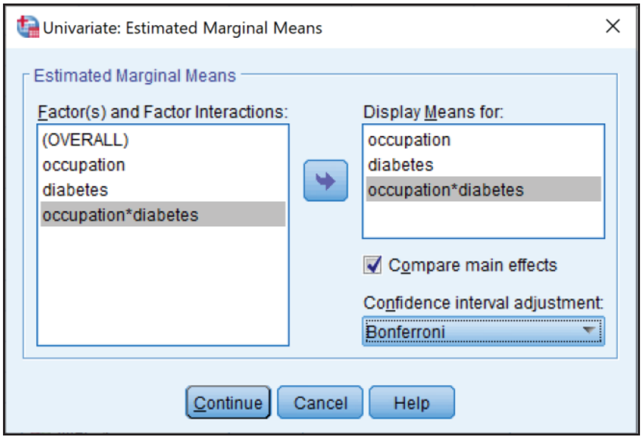
Fixed Factor(s): occupation, diabetes

Random Factor(s):

Covariate(s): age

WLS Weight:

Figure 20.6


Univariate: Estimated Marginal Means

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)
occupation
diabetes
occupation*diabetes

Display Means for:
occupation
diabetes
occupation*diabetes

☑ Compare main effects

Confidence interval adjustment: Bonferroni

20.2.2 Outputs

Table 20.9 Between-Subjects Factors

		Value Label	N
Occupation	1	GOVT JOB	60
	2	PRIVATE JOB	49
	3	BUSINESS	49
	4	OTHERS	52
Have diabetes mellitus	1	Yes	45
	2	No	165

Table 20.10 Descriptive Statistics (unadjusted means)

Dependent Variable: Diastolic BP

Occupation	Have diabetes mellitus	Mean	Std. Deviation	N
GOVT JOB	Yes	82.60	11.559	10
	No	84.18	11.851	50
	Total	83.92	11.720	60
PRIVATE JOB	Yes	79.75	9.468	8
	No	82.80	14.128	41
	Total	82.31	13.443	49
BUSINESS	Yes	84.31	12.216	13
	No	83.19	10.833	36
	Total	83.49	11.096	49
OTHERS	Yes	82.43	7.822	14
	No	80.74	11.733	38
	Total	81.19	10.772	52
Total	Yes	82.53	10.135	45
	No	82.83	12.180	165
	Total	82.77	11.749	210

Table 20.11 Levene's Test of Equality of Error Variances[a]

Dependent Variable: Diastolic BP

F	df1	df2	Sig.
1.284	7	202	.260

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + age + occupation + diabetes + occupation * diabetes

Table 20.12 Tests of Between-Subjects Effects

Dependent Variable: Diastolic BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	384.063[a]	8	48.008	.339	.950
Intercept	99362.441	1	99362.441	701.567	.000
age	15.420	1	15.420	.109	.742
occupation	161.421	3	53.807	.380	.768
diabetes	5.419	1	5.419	.038	.845
occupation * diabetes	126.776	3	42.259	.298	.827
Error	28467.504	201	141.629		
Total	1467419.000	210			
Corrected Total	28851.567	209			

a. R Squared = .013 (Adjusted R Squared = -.026)

Table 20.13 Estimates (adjusted means of diastolic BP) for occupational groups)

Dependent Variable: Diastolic BP

Occupation	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
GOVT JOB	83.400[a]	2.062	79.335	87.465
PRIVATE JOB	81.261[a]	2.300	76.725	85.797
BUSINESS	83.779[a]	1.927	79.979	87.579
OTHERS	81.622[a]	1.864	77.946	85.298

a. Covariates appearing in the model are evaluated at the following values: Age = 26.5143.

Table 20.14 Estimates (adjusted means of diastolic BP) for diabetes

Dependent Variable: Diastolic BP

Have diabetes mellitus	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Yes	82.315[a]	1.823	78.721	85.909
No	82.716[a]	.934	80.874	84.559

a. Covariates appearing in the model are evaluated at the following values: Age = 26.5143.

Table 20.15 Estimates (Occupation * Have diabetes mellitus) [adjusted means of diastolic BP for occupation and diabetes]

Dependent Variable: Diastolic BP

Occupation	Have diabetes mellitus	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
GOVT JOB	Yes	82.643[a]	3.766	75.218	90.069
	No	84.157[a]	1.684	80.835	87.478
PRIVATE JOB	Yes	79.736[a]	4.208	71.439	88.033
	No	82.786[a]	1.859	79.120	86.453
BUSINESS	Yes	84.359[a]	3.304	77.843	90.875
	No	83.199[a]	1.984	79.288	87.110
OTHERS	Yes	82.522[a]	3.193	76.226	88.818
	No	80.723[a]	1.931	76.915	84.531

a. Covariates appearing in the model are evaluated at the following values: Age = 26.5143.

Table 20.16 Pairwise Comparisons of occupation groups

Dependent Variable: Diastolic BP

(I) Occupation	(J) Occupation	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
GOVT JOB	PRIVATE JOB	2.139	3.089	1.000	-6.093	10.371
	BUSINESS	-.379	2.821	1.000	-7.896	7.138
	OTHERS	1.778	2.778	1.000	-5.625	9.180
PRIVATE JOB	GOVT JOB	-2.139	3.089	1.000	-10.371	6.093
	BUSINESS	-2.518	3.002	1.000	-10.519	5.482
	OTHERS	-.361	2.963	1.000	-8.257	7.534
BUSINESS	GOVT JOB	.379	2.821	1.000	-7.138	7.896
	PRIVATE JOB	2.518	3.002	1.000	-5.482	10.519
	OTHERS	2.157	2.678	1.000	-4.978	9.291
OTHERS	GOVT JOB	-1.778	2.778	1.000	-9.180	5.625
	PRIVATE JOB	.361	2.963	1.000	-7.534	8.257
	BUSINESS	-2.157	2.678	1.000	-9.291	4.978

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Table 20.17 Pairwise Comparisons (between diabetics and non-diabetics)

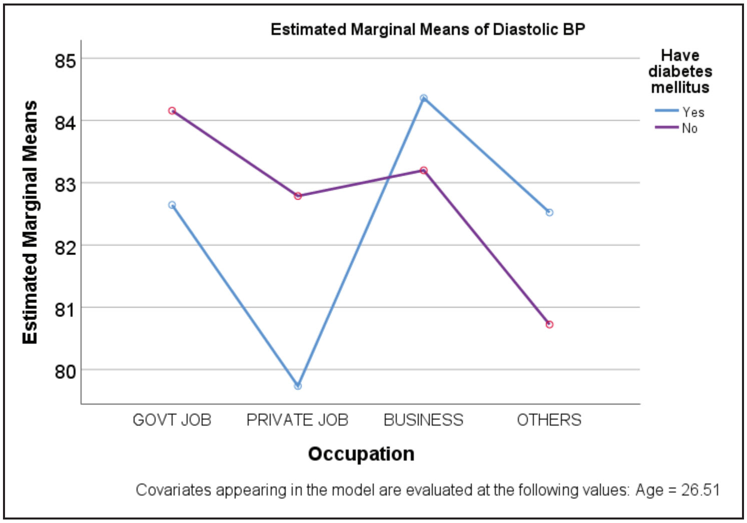
Dependent Variable: Diastolic BP

(I) Have diabetes mellitus	(J) Have diabetes mellitus	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Yes	No	-.401	2.051	.845	-4.445	3.643
No	Yes	.401	2.051	.845	-3.643	4.445

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Figure 20.7 Mean diastolic BP of different occupation groups by diabetes after adjustment for age



20.2.3 Interpretation

Tables 20.9 and 20.10 show the descriptive statistics. All the means provided in Table 22.2 are the crude (unadjusted) means, i.e., without adjusting for age.

Table 20.11 shows the results of Levene's test of Equality of Error Variances. This is the test for homogeneity of variances. We expect the p-value (sig.) to be >0.05 to meet the assumption. In this example, the p-value is 0.260, which is more than 0.05. This means that the variances of the dependent variable (diastolic BP) are similar at each level of the independent variables (occupation and diabetes).

Table 20.12 (tests of between-subjects effects) is the main table showing the results of the two-way ANCOVA test. We tested the hypothesis of whether:

- Mean diastolic BP (in the population) in different occupational groups is same after controlling for age;
- Mean diastolic BP (in the population) among diabetics and non-diabetics is same after controlling for age; and
- Is there any interaction between occupation and diabetes after controlling for age?

Look at the p-values for occupation, diabetes, and occupation*diabetes in Table 20.12. They are 0.768, 0.845 and 0.827, respectively, indicating that none of them are statistically significant. This means that occupation and diabetes do not have any influence on diastolic BP after controlling for age. There is also no interaction between occupation and diabetes after controlling for age. However, we should always check the p-value of the interaction first. If the interaction is significant (p-value <0.05), then the main effects (of occupation and diabetes) are not important because the effect of one independent variable is dependent on the level of the other independent variable.

We can also have information about the influence of the covariate (age) on the dependent variable (diastolic BP). We can see (Table 20.12) that the p-value for age is 0.742, which is not statistically significant. This indicates that there is no significant association between age and diastolic BP after controlling for occupation and diabetes.

Tables 20.13 (estimates for occupation) and 20.14 (estimates for diabetes) show the adjusted means of diastolic BP (dependent variable) at different levels of the independent variables (occupation and diabetes) after controlling for age. In this example, the adjusted mean of diastolic BP of government job holders is 83.4 mmHg (Table 20.13) and that of the diabetics (diabetes mellitus: yes) is 82.3 mmHg (Table 20.14) after controlling for age. Similarly, Table 20.15 shows the adjusted mean of diastolic BP of different occupational groups by diabetes.

Table 20.16 is the table of pairwise comparisons of mean diastolic BP in different occupational groups. *This table is necessary when the independent variable(s) has more than two levels, and there is a significant association between the dependent and independent variables*. Look at the p-values (Sig.) in Table 20.16. Since all the p-values are >0.05, there is no significant difference in mean diastolic BP among the occupational groups after controlling for age.

Figure 20.7 plotted the mean diastolic BP of different occupational groups disaggregated by diabetes. Finally, from the data, we conclude that the diastolic BP is not influenced (there is no association) by occupation and diabetes after controlling for age.

Annex

Table A.1 Codebook of data file <Data_3.sav>

SPSS variable name	Actual variable name	Variable code
ID_no	Identification number	Actual value
age	Age in years	Actual value

sex	Sex: string	m= Male
		f= Female

sex_1	Sex: numeric	0= Female 1= Male
religion	Religion	1= Islam 2= Hindu 3= Others
religion_2	Religion 2	1= Islam 2= Hindu 3= Christian 4= Buddha
occupation	Occupation	1= Government job 2= Private job 3= Business 4= Others
income	Monthly family income in Tk.	Actual value
sbp	Systolic blood pressure in mmHg	Actual value
dbp	Diastolic blood pressure in mmHg	Actual value
f_history	Family history of diabetes	0= No 1= Yes
pepticulcer	Have peptic ulcer	1= Yes 2= No
diabetes	Have diabetes mellitus	1= Yes 2= No
post_test	Post-test score	Actual value
pre_test	Pre-test score	Actual value
date_ad	Date of hospital admission	Actual date
date_dis	Date of discharge	Actual date

References

1. Altman DG. (1992). Practical Statistics for Medical Research (1st Edition). Chapman & Hill.

2. Anderson M, Nelson A. Data analysis: Simple statistical tests. FOCUS on Field Epidemiology: UNC School of Public Health. North Carolina Centre for Public Health Preparedness; Vol 3(6).

3. Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Medical Research Methodology 2003; 3(21):1-13. http://www.biomedcentral.com/1471-2288/3/21

4. Bergmire-Sweat D, Nelson A, FOCUS Workgroup. Advanced Data Analysis: Methods to Control for Confounding (Matching and Logistic Regression). Focus on Field Epidemiology: UNC School of Public Health. North Carolina Centre for Public Health Preparedness; Volume 4, Issue 1.

5. Chan YH. Biostatistics 103: Qualitative Data – Tests of Independence. Singapore Med J 2003; Vol 44(10):498-503.

6. Chan YH. Biostatistics 104: Correlational Analysis. Singapore Med J 2003; Vol 44(12):614-619.

7. Chan YH. Biostatistics 201: Linear Regression Analysis. Singapore Med J 2004; Vol 45(2):55-61.

8. Chan YH. Biostatistics 202: Logistic regression analysis. Singapore Med J 2004; Vol 45(4):149-153.

9. Chan YH. Biostatistics 203. Survival analysis. Singapore Med J 2004; Vol 45(6):249-256.

10. Chan YH. Biostatistics 3.5. Multinominal logistic regression. Singapore Med J 2005; 46(6):259-268.

11. Daniel WW. (1999). Biostatistics: A Foundation for Analysis in the Health Science (7th Edition). John Wiley & Sons, Inc.

12. Field A. (2002). Discovering Statistics Using SPSS for Windows. SAGE Publications: London, California, New Delhi.

13. Gordis L. (2014). Epidemiology (5th Edution). ELSEVIER Sounders.

14. Katz MH. (2011). Multivariable Analysis: A Practical Guide for Clinicians and

Public Health Researchers (3rd Edition). London, Cambridge University Press.

15. Katz MH. (2009). Study Design and Statistical Analysis: A Practical Guide for Clinicians. Cambridge University Press.

16. Katz MH. (2010).Evaluating Clinical and Public Health Interventions – A Practical Guide to Study Design and Statistics. Cambridge University Press.

17. Katz MH. Multivariable Analysis: A Primer for Readers of Medical Research. Ann Intern Med 2003; 138:644–650.

18. Khamis H. Measures of Association: How to Choose? JDMS 2008; 24:155–162.

19. Lee J, Chia KS. Estimation of prevalence rate ratios for cross sectional data: an example in occupational epidemiology. British Journal of Industrial Medicine 1993; 50:861-864

20. Pallant J. (2007). SPSS Survival Manual (3rd Edition). Open University Press.

21. Reboldi G, Angeli F, Verdecchia P. Multivariable Analysis in Cerebrovascular Research: Practical Notes for the Clinician. Cerebrovasc Dis 2013; 35:187–193. DOI: 10.1159/000345491.

22. Szklo M, Nieto FJ. (2007). Epidemiology: Beyond the Basics (2nd Edition). Jones and Bartlett Publishers.

23. Schlesselman JJ, Stolley PD. (1982). Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, Oxford, New York.

24. Tabachnik BG, Fidell LS. (2007). Using multivariate statistics (5th Edition). Boston: Pearson Education.

25. Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med 1998; 55:272–277.

About the Authors

Mohammad Tajul Islam is a Professor (Adjunct) at the North South University and State University of Bangladesh. He teaches epidemiology, data analysis and statistical methods in health sciences for more than 15 years. He is a medical graduate with post-graduation in Tropical Medicine and Epidemiology from the Mahidol University, Bangkok, Thailand. His research interest is maternal and child health. He has authored or co-authored a number of articles published in the international peer reviewed journals. He was involved in a significant number of research projects implemented in Bangladesh, and served as a member of the technical committee for Bangladesh Demographic and Health Survey (BDHS). He is a regular reviewer of the International Journal of Gynaecology and Obstetrics (IJGO) and occasionally reviews articles from other international peer-reviewed journals. He has worked at several UN Agencies (WHO, UNICEF & UNFPA) and international development organizations (Japan International Cooperation Agency and Save the Children) including ICDDR,B for more than 20 years.

Dr. Russell Kabir is working as a Senior Lecturer in Research Methods and Course Leader of MSc Public Health. He has been working as an academic in higher education institutes more than 11 years. He has authored /co-authored in more than 60 journal articles and book chapters. He is currently serving as an Academic Editor for PLOS One and BMC Public Health. Dr. Kabir is interested to perform collaborative and interdisciplinary research in public health issues with a special focus on oral health, reproductive health issues, violence against women and ageing related research.

Dr. Monjura Nisha is an early-career public health researcher at the University of Sydney, Australia. She earned her PhD (Perinatal Epidemiology) from the Sydney School of Public Health, The University of Sydney in 2020. Using epidemiological methods, her PhD topic focused on investigating modifiable risk factors contributing to adverse perinatal outcomes in resource-poor settings. She completed her Master of Public Health (MPH) degree majoring in Epidemiology from 2010-2012, at North South University, Bangladesh. She is adept in the analysis of population based large datasets such as Demographic and Health Survey datasets and applying those to investigate public health related problems.